THESE DE DOCTORAT

De l'Université Paris 7 - Denis Diderot

Spécialité

Chimie Informatique et Théorique

Diversité Moléculaire : Application au Criblage Virtuel, Corrélation avec des Propriétés Physico-chimiques

Soutenue le : 19 septembre 2006

Par : Ana MALDONADO

Devant le jury composé de :

- Prof. Michel DELAMAR (Président)
- Prof. Alexandre VARNEK (Rapporteur)
- Dr. Dragos HORVATH (Rapporteur)
- Dr. Michel PETITJEAN (Examinateur)
- Prof. Georges DIVE (Examinateur)
- Prof. Bo Tao FAN (Directeur de thèse)
- Prof. Jean-Pierre DOUCET (Invité)

This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivs 2.0 License.

A ma famille,
(au sens large du terme)
qui m'a toujours encouragée et supportée,
à ceux et à celles, qui ont cru en moi, je leur dédie ce travail.

Le travail presenté dans ce mémoire a été effectué à l'Institut de Topologie et de Dynamique des Systèmes (ITODYS) de l'Université Paris VII, sous la direction du Professeur Bo-Tao Fan et du Dr. Michel Petitjean. Je les prie de trouver ici mes remerciements les plus sincères pour toute l'aide qu'ils m'ont apportée.

J'exprime également ma profonde gratitude au Directeur du Laboratoire ITODYS, Monsieur le Professeur Michel Delamar pour son accueil chaleureux dans le laboratoire, ainsi que à Madame la Professeur Annick Panaye pour m'avoir acceptée au sein du groupe de Chimie Informatique et pour m'avoir toujours aidée et soutenue tout au long de cette thèse.

Je remercie chaleureusement Monsieur le Professeur Jean Pierre Doucet, pour tous les conseils précieux qu'il n'a jamais cessé de me prodiguer tout au long de ce travail, et dans l'élaboration finale de cette thèse.

Je tiens à exprimer toute ma reconnaissance à Monsieur le Professeur Alexandre Varnek et au Dr. Dragos Horvarth de m'avoir fait l'honneur d' être les rapporteurs de ce mémoire ; Ainsi qu' à Monsieur le Professeur Georges Dive de l'avoir examiné.

J'exprime aussi toute ma gratitude à mes collègues du laboratoire qui m'ont aidée au cours de cette thèse, particulièrement lors de mes nombreux déplacements hors des frontières. Merci d'avoir toujours reçu avec le sourire, mes multiples demandes de services. Fabienne, Florent, Cyril, Lina, Catia, Mme. Wang, … et tous ceux qui se reconnaissent en ces lignes, qu'ils veuillent bien trouver ici l'expression de mes remerciements.

Toutes mes pensées vont aussi à ceux qui m'ont encouragée par des gestes d'amitié dont je leur serais toujours reconnaissante. Merci Ines, Alfredo, Orelle, Cyril, Paul, Véronique et tant d'autres… et merci à toi Raphaël, qui a su être le confident, et le fidèle supporter de tous les instants. Merci à toi et à tous.

TABLE DES MATIERES

page

Introduction

1. Les concepts de similarité et de diversité 1
2. Mesures de similarité et de diversité : éléments principaux 2
   2.1 Les descripteurs 2
   2.2 Les indices de similarité 4
   2.3 Le système des poids 4
3. Problématique et aperçu du système développé 5
   3.1 Problématique particulière 5
   3.2 Aperçu des fonctionnalités 6
4. Plan général 8

Chapitre I. Criblage virtuel et méthodes de traitement structural

I.1 Le criblage virtuel de haut débit en chimie informatique 11
   I.1.1 La chimie combinatoire 12
   I.1.2 Le criblage virtuel et le criblage de haut débit 13
I.2 Le criblage virtuel et les approches structurales 15
   I.2.1 Le choix de descripteurs 16
   I.2.2 Les descripteurs structuraux dans les outils de criblage virtuel 17
   I.2.3 La comparaison de descripteurs dans la littérature 23

Chapitre II. Bases de données : représentation et structuration

II.1. Bases de données. Lexique et construction 31
   II.1.1 Lexique et format de molécules 32
   II.1.2 Construction de la base de données de fragments (FragDB) 33
      II.1.2.1 Les atomes génériques 33
      II.1.2.2 L'origine des fragments et des sous-structures 36
      II.1.2.3 Un aperçu des bases de fragments 39
   II.1.3 Construction des bases de données QueryDB et TestDB 42
II.2. Structuration des informations moléculaires et XML 43
   II.2.1 Les langages de marquage 44
      II.2.1.1 Histoire 44
      II.2.1.2 Principes 45
      II.2.1.3 XML pour structurer les informations chimiques 50
   II.2.2 La structuration de la FragDB avec XML 51
      II.2.2.1 Création et remplissage d'un index-XML de fragments 51
      II.2.2.2 Une DTD pour valider l'index-XML 58
   II.2.3 La structuration du QueryDB et du TestDB 59
      II.2.3.1 Transformation des molécules et création du VecteurRepresentatif-XML 59
      II.2.3.2 Une DTD pour valider le VecteurRepresentatif-XML 64
      II.2.3.3 Une DTD pour valider l'indexResult-XML 65
   II.2.4 La représentation des connaissances 66

Chapitre III. Processus de comparaison de structures moléculaires

III.1 Les recherches structurales 73
III.1.1 Algorithmes de superposition des graphes 73
III.1.2 Recherche de similarité pour des structures moléculaires 77
III.2 Reconnaissance des motifs structuraux et création des vecteurs descripteurs 84
III.2.1 Transformation des molécules et génération des vecteurs descripteurs 86
III.2.1.1 Reconnaissance des motifs structuraux 86
III.2.1.2 Génération des vecteurs-descripteurs 96

Chapitre IV. Mesures de Similarité moléculaires

IV.1 Coefficients et distances 99
IV.2 Comparaisons intermoléculaires 103
IV.2.1 Analyses de Similarité 104
IV.2.2 Calcul de la précision et du rappel « *recall* » 106
IV.3 Les différents niveaux de comparaison 109
IV.3.1 Comparaison exclusivement structurale 112
IV.3.2 Comparaison reposant sur la structure et les propriétés des molécules 116

Chapitre V. Présentation et analyse des résultats

V.1 Analyse de type 1-N 125
V.1.1 Résultats avec la base « Zinc » 126
V.1.2 Résultats avec la base « Random » 135
V.1.3 Comparaison des indices selon le rang 144
V.1.3.1 Graphiques de comparaison d'indices avec la base « Zinc » 144
V.1.3.2 Graphiques de comparaison d'indices avec la base « Random » 147
V.1.4 Comparaison des indices selon la complexité 150
V.1.4.1 Graphiques de comparaison d'indices avec la base « Zinc » 151
V.1.4.2 Graphiques de comparaison d'indices avec la base « Random » 154
V.2 Analyse de type N-N 156
V.2.1 Résultats avec la base « Zinc » 156
V.2.2 Résultats avec la base « Random » 162
V.2.3 Aperçu des résultats structurés et présentés avec XML 166
V.3 Evaluation de l'outil 168
V.3.1 Précision, rappel, et F-measure, pour la base « Zinc » 169
V.3.2 Etude des faux isomorphismes pour des mesures de similarité N-N 173
V.3.3 Limites et avantages de l'outil 175

Chapitre VI. Conclusion et perspectives

VI.1 Conclusions 179
VI.2 Perspectives 184
VI.2.1 Perspectives à moyen terme 184
VI.2.2 Perspectives à long terme 184

Annexes

Annexe 1. Manuel d’utilisation du logiciel 185
Annexe 2. Fichiers XML et structures de données 201
Annexe 3. Format MOL 215
Annexe 4. Tableaux de résultats 221

ABRÉVIATIONS

AAB (Advanced Algorithm Builder): Constructeur avancé d'algorithmes

ADMET (absorption, distribution, metabolism, excretion and toxicity): absorption, distribution, métabolisme, excrétion et toxicité

CAS (Chemical Abstract Service): base de données chimiques de la Société Américaine de Chimie

CML (Chemical Markup Language): Langage de Marquage Chimique

CSS ou SSC (Common Substructure Search): Recherche des Sous-Structures Communes (SSC)

DARC: Description, Acquisition, Restitution, Conception

DISSIM (Statistical module to calculate the DISSIMilarity index): module statistique pour calculer l'index de diversité.

DTD (Document Type Definition): Définition de Type de Document

FREL (Fragments Reduced to an Environment which is Limited): Fragment Réduit à un Environnement Limité

FM (Fragmental Methods): Méthodes fragmentaires

FO (Focus): point de focalisation

GETAWAY (GEometry, Topology and Atom-Weights AssemblY): Assemblage de géométrie, topologie et masses moléculaires

GML (Generalized Markup Language): Langage de Marquage Generalisé

HOMO-LUMO (Highest Occupied Molecular Orbital – Lowest Unoccupied Molecular Orbital): Orbital moléculaire supérieur occupé - orbital moléculaire inférieur non occupé

HTML (Hyper Text Markup Language): Langage de Marquage d'Hyper Texte

HTS (High Throughput Screening): Criblage de Haut Débit

HTSS (Hierarchic Tree Substructure Search Systems): Système de recherche des sous-structures par des arbres hiérarchiques

InkML (Ink Markup Language): Langage de Marquage pour « l'encre digitale »

IR (Infrared): Infrarouge

IUPAC (International Union of Pure and Applied Chemistry): Union International de Chimie Pure et Appliqué.

LaSSI (Latent Semantic Structure Indexing): Indexation structurale sémantique latent

MACCS (Substructure search system from CambridgeSoft Corporation): Système de recherche de sous-structures crée par la corporation CambridgeSoft

MathML (Mathematical Markup Language): Langage de Marquage Mathématique

MDDR (MDL Drug Data Report): Index MDL de données de drogues

MDL (Molecular Design Limited): Corporation vissant au design des nouvelles molécules

MEP (Molecular Electrostatic Potential): Potential electrostatique moléculaire

MCSS ou SSMC (Maximal Common Sub-Structure): Sous-structure maximale commune (SSMC)

Namespace: espace de noms, mot qui permet d'éviter des collisions de noms des balises XML

NP (NP problem): problème NP, c'est-à-dire, que la découverte de l'ensemble des solutions s'effectue en un temps exponentiel

OWL (Web Ontology language): Langage des ontologies du web

QSAR (Quantitative Structure-Activity Relationship): Relation quantitative structure-activité

QSPR (Quantitative Structure-Property Relationship): Relation quantitative propriété-activité

RDF (Radial Distribution Function): Fonction de distribution radiale

RDF (Resource Description Framework): Cadre pour la description des ressources

RuleML (Rule Markup Language): Langage de Marquage de règles

S4 (SubStructure Search Software, Beilstein Institute of Organic Chemistry & Softron Ltd): logiciel de recherche de sous-structures

SGML (Standard Generalized Markup Language): Langage de Marquage Généralisé et Standard

SMD (Standard Molecular Format): Format Moléculaire Standard

SMILES (Simplified Molecular Input Line Entry Specification): Spécification simplifiée de l'entrée linéaire de la molécule

SMIL (Synchronized Multimedia Integration Language): Langage d'intégration multimédia synchronisé

SVM (Support Vector Machines): Moteur de raisonnement vectoriel

SVG (Scalable Vector Graphics): Technique de dessin de vecteurs scalaires

ThermoML (Thermodynamic Markup Language): Langage de Marquage Thermodynamique

UFS (Unsupervised Forward Selection): Sélection non supervisé de descripteurs

UV (UltraViolet): Ultra Violet

VS (Virtual Screening): Criblage Virtuel

W3C (Word Wide Web Consortium): Consortium du WWW

WLN (Wiswesser Line Notation): Notation moléculaire linéaire de Wiswesser

WHIM (Weighted Holistic Invariant Molecular): Descripteur moléculaire par des invariants holistiques

XHTML (Extended HyperText Markup Language): Langage de Marquage Extensible pour HTML

XML (Extended Markup Languages): Langage de Marquage Extensible

XMLSchema (Extended Markup Language Schema): Schéma pour Langage de Marquage Extensible

XQuery (Extended Query): Interrogation des schémas du Langage de Marquage Extensible

XSLT (Extensible Stylesheet Language Transformation): Feuille de style et de transformation pour Langage de Marquage Extensible.

INTRODUCTION

1. Les concepts de similarité et de diversité
2. Mesures de similarité et de diversité : éléments principaux
2.1 Les descripteurs
2.2 Les indices de similarité
2.3 Le système des poids
3. Problématique et aperçu du système développé
3.1 Problématique particulière
3.2 Aperçu des fonctionnalités
4. Plan général

INTRODUCTION

Afin d'identifier de nouvelles molécules susceptibles de devenir des médicaments, la recherche pharmaceutique a de plus en plus recours à des technologies permettant de synthétiser un très grand nombre de molécules simultanément et de tester leur action sur une cible thérapeutique donnée. De récentes évolutions concernent la création d'outils informatiques adaptés au haut débit pour le criblage *in silico* de bases de données chimiques réelles et virtuelles. Le criblage virtuel sert ainsi à réduire des bases qui contiennent un nombre trop important de composants en un ensemble d'éléments prometteurs, par rapport à une cible (ou une famille de cibles) à travers l'application de méthodes informatiques. Une des techniques de criblage virtuel les plus souvent utilisées est sans doute l'analyse de la similarité et de la diversité moléculaire.

1. Les concepts de similarité et de diversité

Avant de continuer, il est important de définir les concepts de similarité et de diversité. Souvent décries comme des concepts flous, la similarité et la diversité ouvrent la porte à une des principales capacités de la logique humaine: **comparer**.

C'est un fait que les animaux et les hommes ont la capacité de distinguer relativement bien les différences entre deux objets. Naturellement, ils font usage de leur sens logique pour reconnaître, regrouper ou généraliser des objets et des concepts vis-à-vis d'une échelle particulière [Rouvray1990]. Il est bien naturel de dire que deux objets sont dissimilaires ou similaires, mais souvent, ceci est relatif à un concept ou plus exactement à une référence établie.

Pour illustrer cette idée, nous pouvons dire que deux animaux sont toujours plus similaires entre eux, qu'un animal et une plante, mais un chat est certainement plus similaire à un tigre qu'à un poisson.

Comme nous l'avons indiqué ci dessus, le concept de similarité structure-propriété a été introduit en chimie pharmaceutique vers le début du 20^{ième} siècle. Mais bien avant, en Grèce antique, la

« méthode scientifique » d'Aristote considérait déjà l'observation et la comparaison comme des étapes indispensables à la découverte de la vérité.

En 1869, Dmitri Mendeleïev propose l'arrangement des éléments chimiques d'une manière périodique [MendeleïevWeb] sur la base de la similarité des propriétés des éléments. Une curieuse observation de sa table fait remarquer des espaces vides qui représentaient des éléments encore inconnus, mais qui avaient été **prédits** par Mendeleïev.

Aujourd'hui, en chimie informatique, le concept de « similarité moléculaire » fournit une méthode simple et populaire pour effectuer du criblage virtuel dans les bases de données chimiques. Elle se sert alors des méthodes de traitement de données comme le groupage (*clustering*) et la fouille de données (*data mining*). D'autre part, la « diversité moléculaire » explore la manière dont les molécules couvrent un espace chimique déterminé à travers la sélection des composants et la construction de bibliothèques combinatoires. Les mesures de similarité et de diversité moléculaires sont donc complémentaires.

2. Mesures de similarité et de diversité : éléments principaux

Pour effectuer des mesures de similarité ou de diversité moléculaire dans un cadre d'analyse moléculaire ou chimique, nous devons prendre en compte trois éléments principaux : les descripteurs, les coefficients et un système de poids.

2.1 Les descripteurs

Les descripteurs sont utilisés pour caractériser les molécules à analyser (voir figure 1). Ils peuvent être calculés à partir de la structure (constitution, configuration et conformation moléculaires) ou des propriétés (physiques, chimiques, biologiques) appartenant aux molécules [Brown1997, Todeschini2000].

Les descripteurs constitutionnels incluent l'information d'ordre des atomes et des liaisons ainsi que la présence ou l'absence de fragments et d'autres caractéristiques 2D. Les descripteurs

configurationnels concernent l'arrangement en 3D des atomes et les descripteurs conformationnels représentent l'arrangement spatial thermodynamique stable des atomes dans une molécule.

Idéalement, les descripteurs utilisés pour le développement des modèles moléculaires devraient être rapidement calculables et facilement interprétables par les ordinateurs et les usagers. Ils devraient représenter la réalité chimique du système et optimiser ainsi la structuration de l'espace chimique [Martin1998].

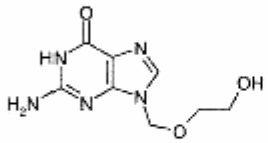
	Représentation Type	Descripteurs Type
1D	C8H10N5O3	Masse moléculaire Nombre d'atomes
2D		Fragments Indices Topologiques Connectivité
3D		Surface Moléculaire Volume Moléculaire Energie d'interaction

Figure 1: Quelques exemples de descripteurs et leur classification en 1D, 2D et 3D.

Les descripteurs moléculaires ont augmenté dernièrement, en nombre et en complexité. La plupart sont obtenus, soit à travers des définitions spécifiques, soit par des combinaisons d'autres descripteurs. Souvent, ils sont composés de valeurs numériques qui correspondent généralement à des propriétés physicochimiques. On compte à ce jour des centaines de descripteurs topologiques, topographiques et de chimie quantique [Katrizky1996]. Dans leur page web, R. Todeschini et V. Consonni [TodeschiniWeb] maintiennent un compteur du nombre de descripteurs moléculaires, et à ce jour, ce nombre atteint 3100.

2.2 Les indices de similarité

Pour mesurer la (dis)similarité moléculaire on utilise des fonctions qui transforment les différences entre une paire de molécules en nombre réels, généralement dans l'intervalle unité [0-1]. Cette quantité fournit une mesure quantitative du niveau de ressemblance chimique [Willett1987, Willett1998].

Les mesures de similarité sont généralement constituées de deux éléments : une représentation mathématique de l'information chimique pertinente (en forme de groupes, graphes, vecteurs ou fonctions) et un index compatible avec la représentation.

Nous allons représenter une molécule M_i sous la forme d'un vecteur où chaque composante *i* correspond à un descripteur moléculaire individuel *di*. D'un point de vue formel, ce vecteur positionne la molécule *M* dans un point de l'espace vectoriel ***V***, dans lequel chacun des axes correspond à un descripteur (figure 2). Cet espace vectoriel s'appelle « l'espace structural » [Maggiora2004].

La (dis)similarité moléculaire entre deux molécules (M_1, M_2) sera intuitivement reliée à la distance entre les deux points dans cet espace particulier. La règle de calcul de cette distance est appelée « métrique ».

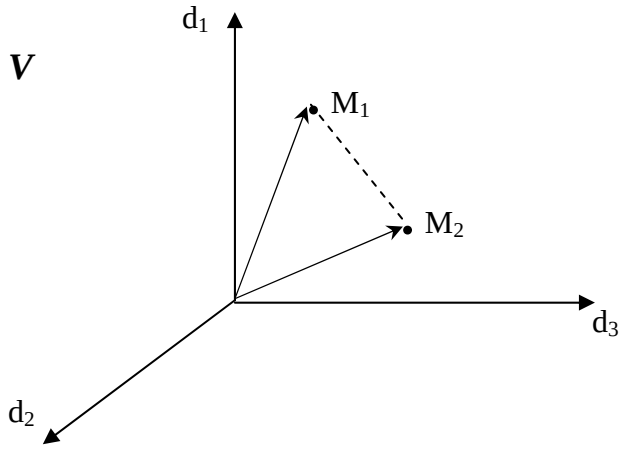


Figure 2. L'espace structural de deux molécules représentées par des descripteurs d_1, d_2 et d_3

Ainsi, toute mesure adéquate de la similarité doit être cohérente avec les propriétés d'une distance mathématique [Petitjean1996].

L'évaluation de similarité peut être abordée par des corrélations, des mesures de distance ou des approches probabilistes ou associatives. La performance de différentes mesures de similarité est le sujet de nombreux travaux [Pearlman1999, Willet1986, Holliday2002].

Remarquons que l'évaluation de similarité se fait dans l'espace structural défini par les descripteurs choisis au moyen d'une métrique fixée et non par rapport aux distances interatomiques dans l'espace 3D.

2.3 Le système des poids

Le troisième élément est le système de poids, qui est utilisé pour assigner différents niveaux d'importance aux différents composants d'une représentation. Il y a des travaux intéressants sur la manière qu'ont les poids d'exercer une influence sur l'utilité de la mesure de similarité moléculaire [Bath1993, Sadowski1998]. Comme notre intérêt est centré sur la chimie informatique et ses applications, le critère adopté pour notre étude sera en relation avec la chimie médicinale et pharmaceutique. En conséquence, la diversité moléculaire pourra être exprimée comme la différence de propriétés physicochimiques et de structure inhérente à chaque molécule.

Trouver une définition satisfaisante pour nos besoins, mais suffisamment générale, aux concepts de similarité et de diversité est très difficile. Des approches différentes pourront être adaptées pour des critères particuliers et permettre ainsi de trouver des solutions à des problèmes ponctuels. En tout cas, les informations chimiques ainsi que les critères de similarité ne devront avoir aucune ambiguïté, notamment pour le traitement informatique du problème.

3. Problématique et aperçu du système développé

3.1 Problématique particulière

Dû au nombre élevé des techniques de criblage virtuel et de haut débit, nous sommes obligés de circonscrire notre problème à un cadre plus succinct. Nous avons déjà indiqué dans la section 1 de

cette introduction, les problèmes liés au traitement des bases de données chimiques : l'augmentation de leur taille, mais aussi le souci de diversité qui règne aujourd'hui dans les centres de criblage. La recherche de nouvelles molécules semble donc être au cœur des besoins actuels dans beaucoup de domaines liés à la chimie industrielle, organique, médicinale, etc.

Notre problématique est centrée sur l'analyse de grandes bases de données chimiques. Notre méthode consiste à effectuer des analyses de similarité et de diversité en utilisant une approche mixte structure-propriétés, pour comparer des molécules ou des bases de molécules et extraire des connaissances utiles au criblage, à l'analyse et à l'amélioration de ces mêmes bases moléculaires. Pour implémenter notre méthode, nous nous sommes vus dans le besoin de structurer les informations chimiques contenues dans les bases moléculaires en utilisant des langages de marquage et de construire un outil qui effectue de manière automatique les analyses sur les bases de molécules.

3.2 Aperçu des fonctionnalités

Nous avons donc mis au point une base de fragments qui est à l'origine de l'approche structure-propriété qui caractérise notre outil. Les informations chimiques des sous-structures contenues dans la base seront codées et utilisées ensuite pour construire des descripteurs moléculaires. Les descripteurs moléculaires ainsi construits coderont l'information structurale et physicochimique de la molécule cible.

Dans notre exemple (figure 3), la structure moléculaire (la 1-(3-amino-cyclohexyl)-ethanone) est analysée pour générer un vecteur, afin de mettre en évidence la présence ou l'absence de certaines sous-structures (fragments) prédéfinies et référencées dans une base des sous-structures « de référence » (que nous appellerons par la suite FragDB).

Chaque sous-structure de référence (ici : CNUQ6-074bs, AGCC-014Q et ANSZ-000Z) est associée à un élément du vecteur. Ces éléments pointent vers des informations de nature diverse.

Si la molécule M contient la sous-structure de référence, cette structure sera prise en compte pour la construction du vecteur.

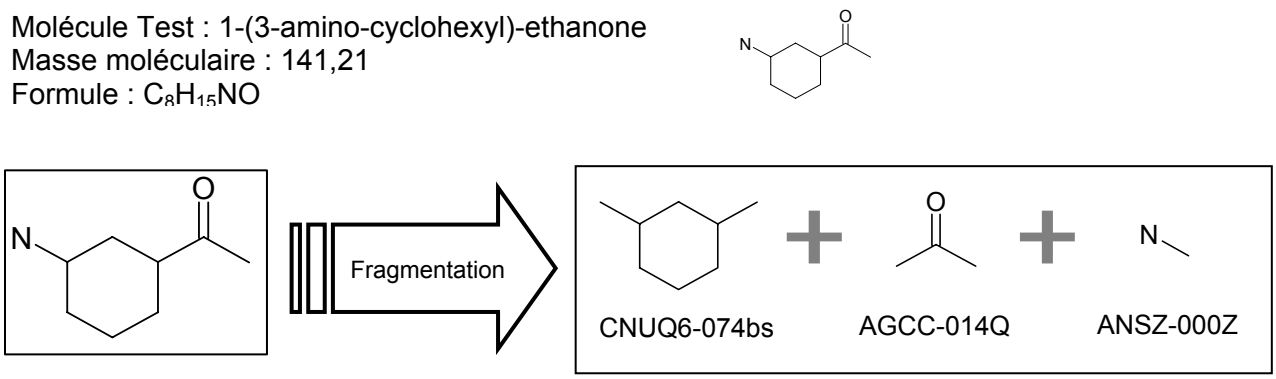


Figure 3. Analyse d'une molécule en utilisant des sous-structures pour sa description.

Une fois les vecteurs descripteurs construits, une mesure de distance est établie entre les molécules appartenant à la base de molécules requête (QueryDB) et celles de la base de molécules test (TestDB). Différentes formules de calcul de la similarité sont choisies dans une liste de possibilités pour adapter le calcul aux besoins divers de l'usager. Quatre types de comparaisons différentes (1-1 à N-M) sont proposés donnant des informations à interprétation diverse. Trois niveaux de complexité (par un système de poids) ont également été implémentés. Les mesures de similarité ou diversité ainsi obtenues pourront être interprétées par l'usager pour trouver des molécules ayant les propriétés ou la structure de la molécule cible, pour rendre plus hétérogène une base de molécules, optimiser une base de réactifs, etc.

Dans la figure suivante les parties principales de l'outil de criblage virtuel proposé, ainsi que la procédure suivie pour le calcul de similarités sont présentées. On peut distinguer cinq étapes principales:

1) Choix de la QueryDB et de la TestDB à partir des fichiers .mol fournis par l'usager,

2) Génération pour chaque fichier .mol d'un vecteur descripteur correspondant que l'on indexe,

3) Choix des types de comparaisons (1-1 à N-M),

4) Choix des niveaux de complexité (poids propriétés et/ou poids sous-structure),

5) Choix des formules et calcul de la similarité.

Les résultats de l'analyse sont détaillés sous forme de tableaux et de graphiques, pour augmenter ainsi les interprétations possibles des résultats. Une représentation alternative consiste à établir un classement (*ranking)* des molécules, selon leur score de (dis)similarité. Ainsi, à la fin de la procédure, les molécules de la TestDB sont ordonnées selon leur degré de ressemblance avec la (ou les) cible(s). Cette méthode facilite l'élaboration de listes de molécules potentiellement intéressantes selon les critères choisis, molécules à tester ou à synthétiser en priorité. De cette manière, on fait des économies dans le processus de recherche de nouvelles drogues ou de molécules actives.

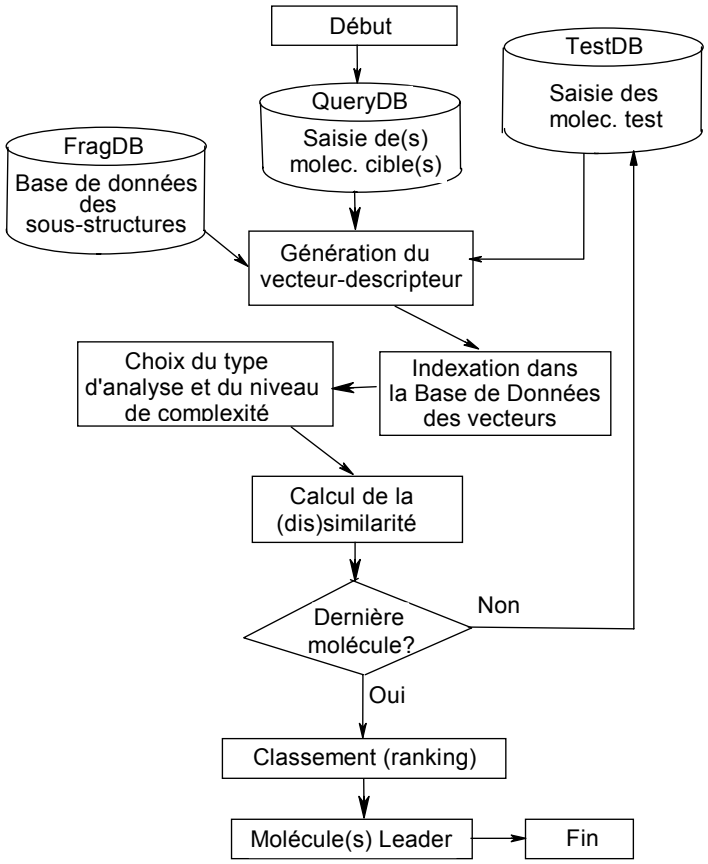


Figure 4. Aperçu des fonctionnalités du système développé

4. Plan général

Cette thèse est consacrée à la réalisation d'un système de calcul de similarité et de diversité à partir de descripteurs structuraux et des propriétés physicochimiques. Le mémoire comprend les parties suivantes :

Le chapitre 1 contient une présentation des méthodes et d'outils de traitement structural 2D apparus dans la littérature, ainsi que l'état de l'art des techniques de criblage virtuel en chimie informatique.

Dans le chapitre 2, l'approche structurale utilisée pour représenter les fragments et les molécules sera expliquée. La construction de cette représentation ainsi que sa syntaxe et sa structuration utilisent un langage de marquage (XML). Le chapitre 3 concerne la création des vecteurs moléculaires, les recherches structurales et les méthodes de comparaison de descripteurs. Le chapitre 4 traite des critères pour effectuer les mesures de similarité et de diversité moléculaire dans notre approche. Les résultats de nos calculs seront présentés et analysés dans le chapitre 5. Des applications possibles en gestion des bases de données et en chimie médicinale seront envisagées. Nous finirons ce manuscrit avec la conclusion et les perspectives futures.

Références

[Bath1993] Bath, P.A., Morris, C.A.,Willett, P., *Effects of Standardization on Fragment-Based Measures of Structural Similarity*, J. Chemomet., 7 (1993) 543-550.

[Brown1997] Brown, R.D., *Descriptors for diversity analysis*, Persp. Drug Disc.Design, 7/8 (1997) 31-49.

[Holliday2002] Holliday, J.D., Hu, C.Y., Willett, P., *Grouping of coefficients for the calculation of Inter-molecular similarity and dissimilarity using 2D fragment Bit-Strings,* Comb. Chem. High Throughput Screening, 5 (2002) 155-166.

[Katrizky1996] Katritzky, A.R., Lobanov, V.S., Karelson, M., *CODESSA Reference Manual, Version 2.0*, Gainville, 1996.

[Maggiora2004] Maggiora, G.M., Shanmugasundaram, V., *Molecular Similarity Measures*. In Methods in Molecular Biology, vol. 275. Chemoinformatics. Concepts, Methods and Tools for Drug Discovery. Bajorath, J. (Ed.) Humana Press Inc., Totowa, NJ. 2004. pp.1-50.

[Martin1998] Martin, Y.C., Bures, M.G., Brown, R.D., *Validated Descriptors for Diversity Measurements and Optimization*, Pharm. Pharmacol. Commun., 4 (1998) 147-152.

[Mendeleïev Web] Information disponible à: http://pearl1.lanl.gov/periodic/mendeleev.htm

[Pearlman1999] Pearlman, R.S., *Novel Software Tools for addressing Chemical Diversity*, Network Science (1999). Disponible à: http://www.netsci.org/Science/Combichem/feature08.html

[Petitjean1996] Petitjean, M., *Three-Dimensional Pattern Recognition from Molecular Distance Minimization*, J. Chem. Inf. Comput. Sci., 36 (1996) 1038-1049.

[Rouvray1990] Rouvray, D.H., *The evolution of the concept of molecular similarity*. In Johnson, M.A., Maggiora, G.M. (Eds.) Concepts and Applications of Molecular Similarity, John Willey & Sons, New York, 1990. pp. 15-42.

[Sadowski1998] Sadowski, J., Kubinyi, H., *A Scoring scheme for discriminating between drugs and non drugs*, J. Med. Chem., 41 (1998) 3325-3329.

[Todeschini2000] Todeschini, R., Consonni, V., *Handbook of Molecular Descriptors*, In Mannhold, R., Kubinyi, H.,Timmerman, H. (Eds.) Series of Methods and Principles of Medicinal Chemistry - vol. 11, Wiley-VCH, New York, 2000.

[TodeschiniWeb] Information disponible à: http://www.disat.unimib.it/chm/QSARnews2.htm

[Willet1986] Willett, P., Winterman, V.A. *Comparison of some measures for the determination of intermolecular structural similarity measures*, Quant. Struct. -Act. Relat., 5 (1986) 18-25.

[Willett1987] Willett, P. (Ed.) Similarity and clustering in chemical information systems, Research Studies Press, Letchworth, Herts., U.K., 1987.

[Willett1998] Willett, P., Barnard, J.M., Downs, G.M., *Chemical Similarity Searching*, J. Chem. Inf. Comput. Sci., 38 (1998) 983-996.

CHAPITRE I.

CRIBLAGE VIRTUEL ET METHODES DE TRAITEMENT STRUCTURAL

I.1 Le criblage virtuel de haut débit en chimie informatique
I.1.1 La chimie combinatoire
I.1.2 Le criblage virtuel et le criblage de haut débit
I.2 Le criblage virtuel et les approches structurales
I.2.1 Le choix de descripteurs
I.2.2 Les descripteurs structuraux dans les outils de criblage virtuel
I.2.3 La comparaison de descripteurs dans la littérature

CHAPITRE I. CRIBLAGE VIRTUEL ET METHODES DE TRAITEMENT STRUCTURAL

Dans ce chapitre, nous présenterons les concepts et l'histoire des techniques de criblage virtuel et de haut débit. Nous montrerons leur application à l'interrogation des bases de données et à l'analyse de la similarité et de la diversité des molécules. Nous allons expliquer également comment et pourquoi nous avons choisi des descripteurs structuraux au sein de notre outil de criblage virtuel. Nous finirons avec un état de l'art des outils de criblage virtuel qui utilisent des descripteurs structuraux et qui relèvent notre problématique.

I.1 Le criblage virtuel de haut débit en chimie informatique

Enrichir le "panorama chimique" et proposer de nouvelles sources de diversité moléculaire a été depuis longtemps un des buts principaux des chimistes. Ainsi, le « principe de similarité des propriétés » [Johnson1990, Martin2002] qui affirme, depuis une centaine d'années, que « des molécules structuralement similaires auront des propriétés similaires », a servi de source à la découverte de nouvelles molécules, même si ce principe a été mis en cause récemment [Doucet1998, BajorathWeb, Nikolova2003].

L'intérêt pour la diversité moléculaire remonte donc à l'application des concepts de similarité et de diversité en chimie (voir ces définitions dans l'introduction), et plus particulièrement en chimie pharmaceutique. Les premiers travaux sur la relation entre structure et propriétés physicochimiques dans les molécules simples et organiques datent respectivement de 1842 [Kopp1842] et de 1864 [Richardson1875]. On remarquera les recherches de B.W. Richardson, auteur d'une série de travaux scientifiques sur la toxicologie, qui mettaient déjà en évidence les effets nocifs de l'alcool et du tabac. Mais il faut attendre 1947 pour voir apparaître des descripteurs structuraux, et des indices topologiques [Wiener1947]. D'autres sources de diversité moléculaire ont été puisées dans la chimie des peptides. Les combinaisons possibles étant très nombreuses, ceci a rendu plus difficile le travail

de synthèse des chimistes de l'époque. Une nouvelle procédure a alors révolutionné la manière d'aborder ce problème: au lieu de synthétiser des molécules cible, après un long processus de sélection et d'isolation, on a commencé à synthétiser des mélanges de produits, et à tester les propriétés de ceux-ci.

I.1.1 La chimie combinatoire

La chimie combinatoire (réelle ou virtuelle) est apparue naturellement comme une option viable au problème de la diversité moléculaire. Aujourd'hui, c'est un moyen pratique pour prédire et synthétiser une grande quantité de molécules en chimie pharmaceutique et agrochimique [Moos1996, Willett1997, Weber2000]. Comme moteur de diversité, cet outil est devenu indispensable et a joué un rôle important dans le progrès de la synthèse automatique et parallèle, survenu ces vingt dernières années [Stu2003].

Cette méthode repose sur l'idée d'obtenir le plus grand nombre de produits possibles, d'une réaction particulière et ceci sous certaines conditions (voir [Gordon1998] et la figure I.1 pour plus d'exemples). Comme son nom l'indique, ces possibilités dites « combinatoires » ne sont pas infinies, mais très nombreuses, d'où le problème du traitement (réel ou virtuel) de ces molécules.

Aux données combinatoires s'ajoutent de nouvelles molécules, issues des synthèses, des extractions et d'autres procédés chimiques, dans les bases de données chimiques à caractère académique ou industriel. Ainsi, chaque année, le CAS (*Chemical Abstract Service*) voit sa base de molécules chimiques augmenter de millions de nouveaux composants. Les structures, les propriétés physicochimiques et biologiques de ces molécules sont ensuite codées et enregistrées, générant plus d'informations.

L'organisation, l'analyse, la recherche et la gestion de cette grande quantité d'informations ouvre de nouvelles possibilités aux techniques novatrices de chimie informatique, parmi lesquelles on compte le criblage de haut débit -virtuel ou réel- (*virtual screening* et *high troughput screening*), la fouille de données (*data-mining*), etc.

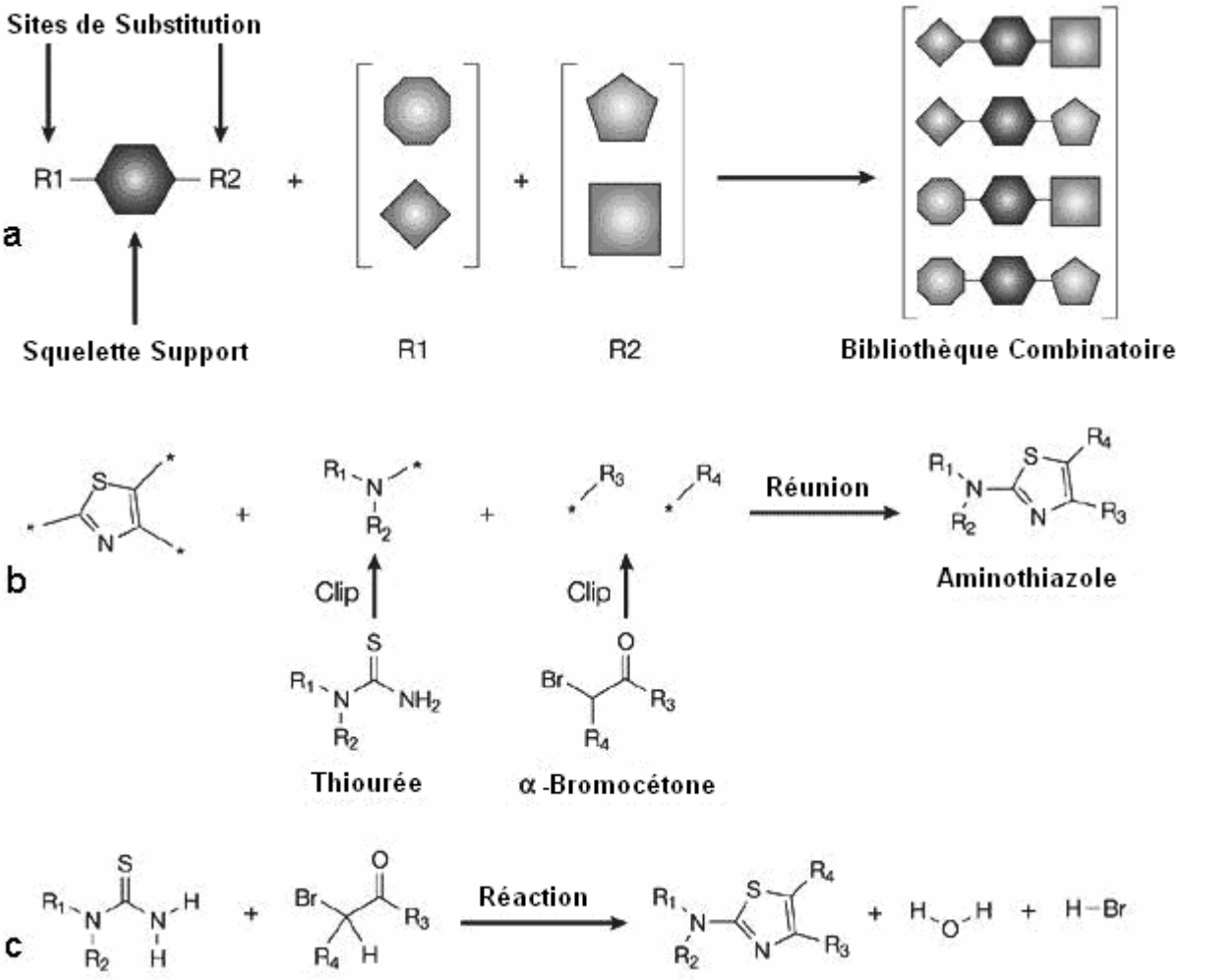


Figure. I.1. Génération d'une bibliothèque virtuelle, où deux approches sont couramment utilisées: (a) La première est basée sur les structures de Markush. (b) La deuxième consiste à attacher systématiquement les réactifs aux sites actifs. (c) Dans une variation de la deuxième approche, des parties spécifiques des réactants sont spécifiées ainsi que la nature des réactions possibles [OFarrell2005].

I.1.2 Le criblage virtuel et le criblage de haut débit

Le criblage virtuel est une technique relativement récente. Ses origines se situent dans les années 70 avec les premiers efforts pour effectuer des recherches 2D avec des fragments structuraux et des cibles 3D, pour ensuite se concentrer dans l'automatisation du *docking* des ligands dans les sites de liaison protéinique. Aujourd'hui le criblage virtuel se divise en une grande diversité de méthodes :

- Approches basées sur la structure du récepteur (*target structure-based VS*),
- Approches basées sur la structure du ligand (*drug-based VS*),
- Approches basées sur des vecteurs structurant des informations chimiques (*fingerprints, pharmacophore, etc.*),
- Techniques de classification des molécules (*cluster analysis, cell-based partitioning)*
- Méthodes statistiques (*3D/4D QSAR models*), etc.

Ainsi, pour aboutir le plus vite possible et à un moindre coût aux molécules désirées, les bases de molécules sont passées au crible [Stahura2004]. Ce criblage doit être réalisé à haut débit pour les bases de données de grande taille, afin d'obtenir des résultats dans un temps raisonnable (voir [Walters1998] et la figure I.2)

Les candidats retenus après le premier criblage peuvent être soumis à d'autres filtres par rapport à des propriétés calculables sur la base de modèles empiriques à partir de leur structure (par exemple, l'affinité pour les graisses, la solubilité...). Ces critères serviront à trier les molécules qui, compte tenu de ces propriétés, ont le plus de chance d'être actives en fonction de telle ou telle cible.

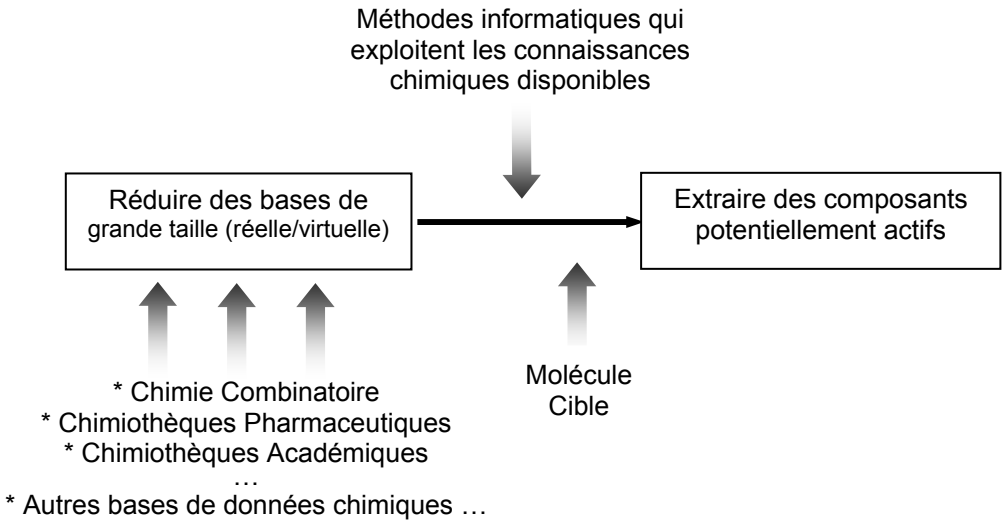

Méthodes informatiques qui exploitent les connaissances chimiques disponibles
Réduire des bases de grande taille (réelle/virtuelle)
Extraire des composants potentiellement actifs
Molécule Cible
* Chimie Combinatoire
* Chimiothèques Pharmaceutiques
* Chimiothèques Académiques
…
* Autres bases de données chimiques …

Figure I.2. Schéma explicatif du processus de criblage virtuel de bases de données chimiques.

Le criblage virtuel et le criblage réel (que ce soit de haut ou de bas débit) sont des techniques complémentaires dans la recherche de nouvelles molécules. Le criblage réel est actuellement le procédé le plus utilisé en chimie médicinale [Bocker2004]. Il consiste à identifier les molécules actives par mise en contact avec la cible biologique. Ces cibles peuvent, par exemple, être des protéines dont on a identifié expérimentalement l'implication dans tel ou tel processus pathologique. Depuis 30 ans, des progrès dans la robotique et l'automatisation ont permis de multiplier les tests et de réduire les coûts car les essais sont « miniaturisés » et utilisent des volumes d'échantillons très réduits. Ces essais reposent sur des systèmes capables de réaliser des taches séquentielles indépendantes telles que dilution, pipetage et répartition de composés dans des puits, agitation, incubation et finalement lecture et analyse de résultats. Ils sont pilotés par des logiciels spécifiquement adaptés au type de tâche à réaliser.

De nombreux travaux décrivent de manière assez complète les méthodes de criblage virtuel (complémentaires ou non au HTS) qui ont été adaptées ou créées pour l'analyse, la classification, la sélection ou le filtrage des bases de données moléculaires. [Stahura2004, Böcker2004, Lengauer2004, Bajorath2002]

I.2 Le criblage virtuel et les approches structurales

Les domaines d'application du criblage virtuel (VS) et du criblage de haut débit (HTS) sont tout à fait différents. Ainsi le VS est souvent discuté dans un contexte chemoinformatique tandis que le HTS appartient au domaine « réel » de la recherche pharmaceutique. Nous ne ferons pas ici une discussion détaillée de toutes les techniques englobées par les termes VS et HTS, car ceci est hors des objectifs de ce manuscrit. Nous expliquerons plutôt, comment et pourquoi nous avons choisi des descripteurs de type structural pour le traitement et l'analyse des molécules. Pour cela, nous dresserons une liste de travaux de comparaison des descripteurs. Leurs conclusions nous mèneront à l'état de l'art des approches structurales utilisées pour les outils de criblage virtuel relevant de notre problématique.

I.2.1 Le choix de descripteurs

Dans l'introduction, nous avons présenté brièvement quels étaient les éléments principaux pour effectuer des mesures de similarité et de diversité dans un cadre moléculaire. Les descripteurs figurent parmi ces éléments. Le calcul et la sélection des descripteurs sont des facteurs déterminants de la réussite du criblage virtuel de molécules. Beaucoup de questions doivent donc être posées. Si des propriétés physicochimiques sont utilisées, il faut fixer à l'avance lesquelles seront retenues et comment elles devront être calculées. Dans le cas de descripteurs structuraux, il faut choisir le niveau de représentation (1D, 2D ou 3D) en sachant que l'approche 1D présente de nombreux avantages, mais est d'un niveau descriptif incomplet; les descripteurs 2D reflètent bien les propriétés physiques et la réactivité dans la plupart des cas, mais l'activité biologique est étroitement liée à la représentation 3D. Cependant, l'utilisation de structures 3D dans la caractérisation des molécules présente des problèmes de conformation, d'énergie et aussi de disponibilité des bases de données 3D. D'autre part, les tautomères et les ions présentent de nouvelles contraintes.

Des approches dites « mixtes » sont très utilisés actuellement, mais là encore il faut choisir un groupe de descripteurs en veillant à leur indépendance et à leur utilité. Dans ce choix, le problème à traiter est souvent NP complet, c'est-à-dire un problème pour lequel le temps de résolution peut s'avérer exponentiel. Ainsi, l'usage de techniques d'apprentissage automatique semble nécessaire. En raison de l'existence de bases de molécules de plus en plus grandes, le facteur de vitesse de traitement ne pourra pas être négligé au moment de choisir la représentation optimale.

Il est important de noter qu'il n'existe pas de « bon » ou de « mauvais » descripteur : l'utilité et l'efficacité sont étroitement liées aux types de molécules à traiter ainsi qu'au calcul à effectuer. Par conséquent, la plupart des descripteurs connus aujourd'hui sont employés de préférence dans le contexte pour lesquels ils ont été créés.

De nombreux travaux abordent les tâches difficiles de calcul, sélection et comparaison des descripteurs. Par exemple, la méthode UFS (*Unsupervised Forward Selection*) de **Whitley**

[Whitley2000] permet de calculer de grandes quantités de descripteurs et d'éliminer ensuite tous ceux qui ont un coefficient de corrélation supérieur à une valeur déterminée.

I.2.2 Les descripteurs structuraux dans les outils de criblage virtuel

La représentation d'une molécule comme une fonction de sa structure ou de ses sous-structures est un moyen communément utilisé pour les chimistes dans la recherche de similarité et la gestion des bases de données chimiques. Traditionnellement ces descripteurs 2D sont liés à la taille et à la connectivité de la molécule, à la présence de groupes fonctionnels, etc. Ces caractéristiques leur donnent une place très importante dans le groupe de descripteurs moléculaires 1D, 2D, 3D.

Précédemment dans l'introduction, nous avons indiqué que les descripteurs utilisés pour le développement des modèles moléculaires devraient représenter la réalité chimique du système, être rapidement calculables et facilement interprétables par les ordinateurs et les usagers. Ceci dit, on compte aujourd'hui avec de multiples représentations moléculaires. Dans la section I.2.3, une sélection des travaux comparatifs de fiabilité et d'efficacité des descripteurs sera présentée.

Il est important de noter que plusieurs de ces travaux coïncident avec l'idée que les descripteurs sous-structuraux présentent un rapport « efficacité-simplicité du modèle » assez avantageux. D'ailleurs ils ont été largement utilisés dans la communauté pour s'attaquer aux problèmes de criblage de bases de données, d'optimisation de bibliothèques, et de prédiction des propriétés entre autres.

Un échantillon des travaux abordant les approches structurales pour des outils de criblage virtuel et d'analyse de la similarité et de la diversité moléculaire est présenté par la suite. Plusieurs approches ont été traitées au même niveau : les approches utilisant des graphes ou des sous-graphes, le calcul d'indices topologiques en utilisant des fragments ou des sous-structures générés automatiquement, et enfin l'analyse de la ressemblance moléculaire à travers les environnements atomiques (atomes, fragments ou liaison autour d'un nœud).

Gillet [Gillet2003] a étudié l'efficacité des graphes pour les recherches de similarité. Elle a

démontré que l'on peut définir une hiérarchie de graphes et que ceux-ci peuvent être utilisés pour trouver des similarités entre composants appartenant à différentes séries chimiques (figure I.3) et aider à l'identification de composants avec la même bioactivité.

Cuissart [Cuissart2002] a utilisé l'extraction de sous-structures des molécules cibles, comme clef de recherche des nouvelles molécules. Il est possible de chercher soit des isomorphismes (i.e. *common substructure/subgraph* (CSS) ou *maximal common substructure/subgraph* (MCSS)) soit des homomorphismes des graphes. La similarité entre les molécules est calculée ensuite en utilisant le nombre calculé d'atomes communs. Ces descripteurs ont montré leur efficacité pour établir des relations structure - dégradation biologique.

Japertas [Japertas2002] a appliqué la « méthode fragmentaire (FM) » pour la recherche de nouveaux composants et pour la prédiction de propriétés physiques et biologiques. Il a proposé un nouveau système appelé *Advanced Algorithm Builder* (AAB), lequel utilise des FM pour construire des modèles QSPR, QSAR et SAR. La figure I.4 illustre comment la fragmentation des structures s'effectue.

Ivanciuc [Ivanciuc2000] explore des nouveaux indices topologiques obtenus à partir du calcul des graphes moléculaires. Dans son travail, l'auteur montre que ceux-ci sont des descripteurs structuraux potentiels pour la caractérisation de la diversité moléculaire.

Randic [Randic1979] propose un ordre théorique des graphes structuraux comme un outil pour effectuer des recherches systématiques de similarité dans des bases de données moléculaires. Dans un autre article [Randic2001], l'auteur introduit un nouveau descripteur moléculaire basé sur le nombre de couches de valence à partir des noeuds d'un graphe moléculaire. Cette approche a été validée en faisant des calculs du point d'ébullition, de l'entropie et de la densité des octanes.

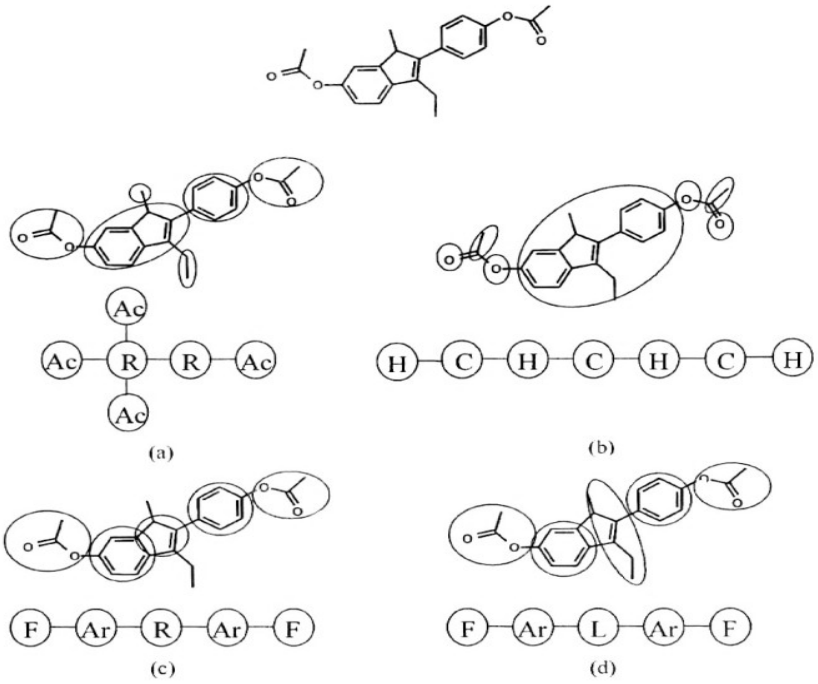


Figure. I.3. Exemples de différents graphes réduits qui peuvent être générés pour les structures montrées. En (a) les noeuds correspondent aux systèmes cycliques (R) et aux éléments acycliques connectés (Ac); En (b) les noeuds correspondent aux éléments carbone (C) et aux éléments hétéroatomiques (H); En (c) les noeuds correspondent aux anneaux aromatiques (Ar), anneaux aliphatiques (R) et groupes fonctionnels (F); En (d) les noeuds correspondent aux anneaux aromatiques (Ar), groupes fonctionnels (F) et groupes de liaison (L).

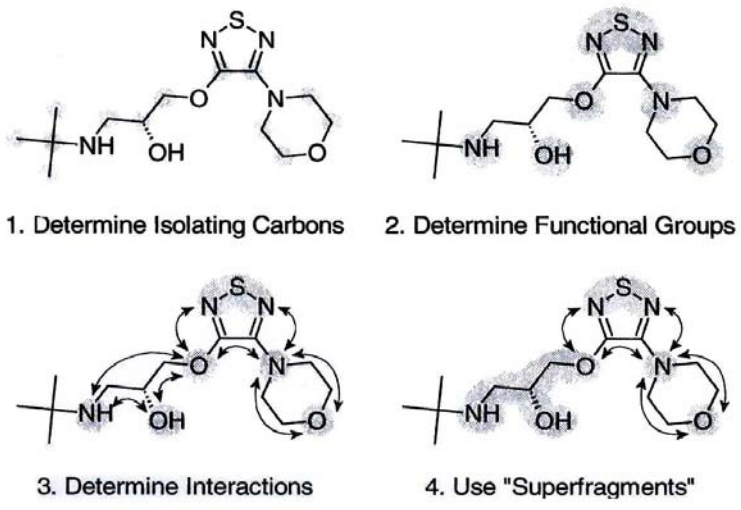


Figure I.4. Fragmentation de structures chimiques complexes (timolol) suivant la méthode fragmentaire (FM).

Environnement moléculaire. La représentation d'une molécule comme fonction de son environnement (atomes, fragments ou liaison autour d'un nœud) est souvent utilisée comme un type de descripteur sous-structural.

Le système DARC développé par **Dubois** [Dubois1986, Dubois1999], décrit les sous-structures contenues dans une molécule à travers le concept de FREL. Les FRELs sont des sous-structures ordonnées d'une manière concentrique autour d'un foyer (FO). Le foyer peut être un atome ou une liaison de la molécule cible, voir figure I.5.

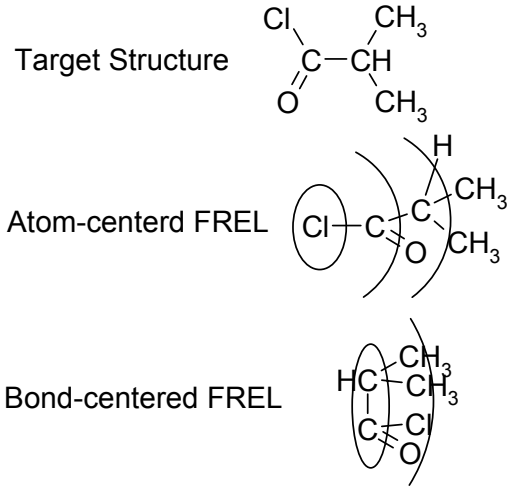


Figure I.5. FREL: Fragment Réduit à un Environnement Limité

La génération des FRELs obéit à certaines étapes : la molécule originale est transformée dans un graphe chromatique équivalent; ensuite, le graphe est focalisé sur la liaison ou l'atome voulu; et à la fin, l'ordre linéaire par rapport au FO est généré. La figure I.6 montre un exemple d'extraction de FREL. Ici, le FO est un groupe hydroxyle et un carbone alpha. L'environnement du FO peut être choisi en accord avec la profondeur désirée de l'analyse.

Cette approche offre l'avantage de pouvoir paramétrer le FO en fonction de la propriété étudiée. On peut également choisir la profondeur de l'environnement, ceci pouvant être généré algorithmiquement d'une manière automatique.

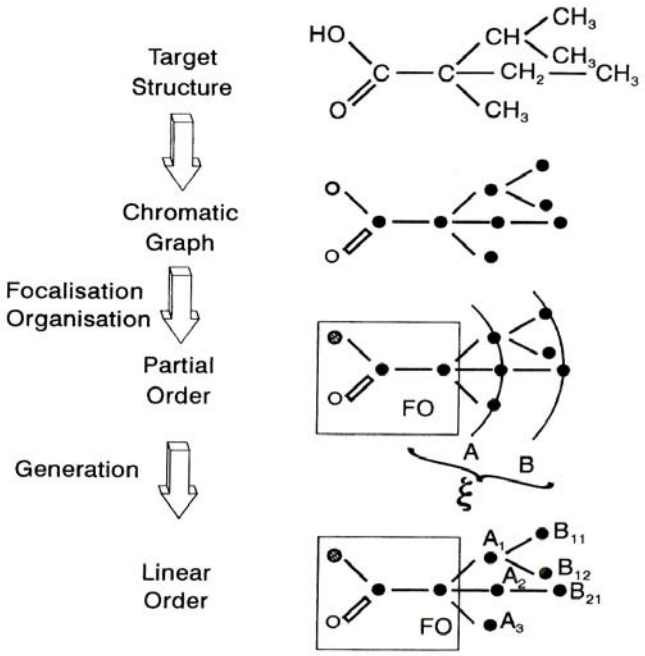


Figure. I.6. Génération d'un ordre linéaire à partir d'une structure cible.

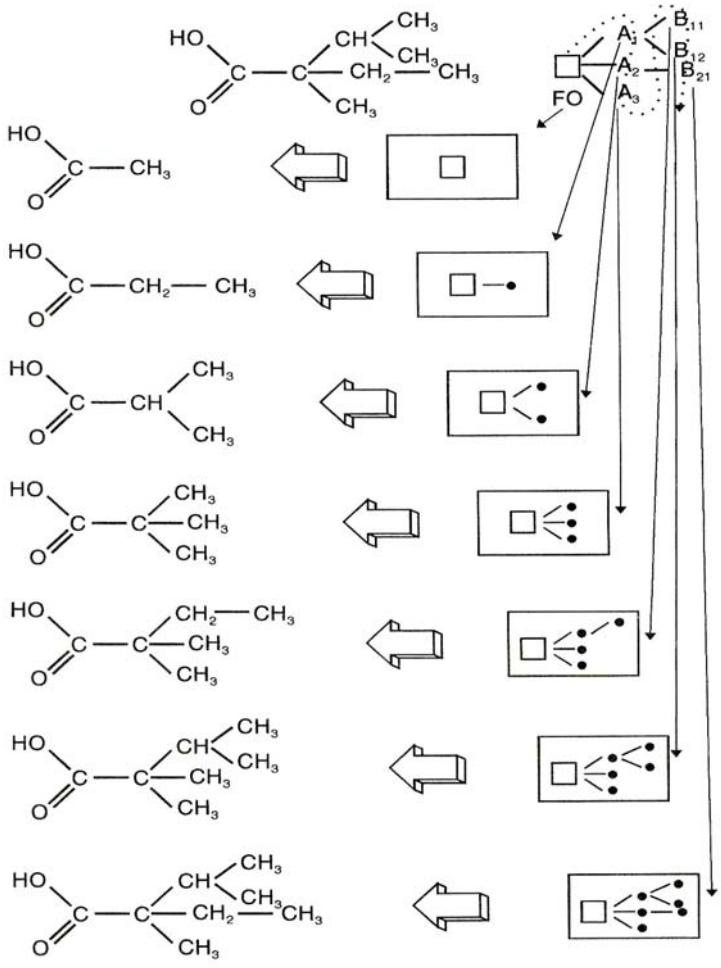


Figure. I.7. Génération de FRELs pour une molécule cible [Dubois1999].

Dans l'approche DARC, la nature des atomes est spécifiée en utilisant des graphes colorés, ce qui simplifie énormément le modèle. La figure I.7 montre un exemple de génération de FRELs à partir d'une molécule cible. L'extraction des FRELs peut être effectuée dans tous les atomes et dans toutes les liaisons.

Dans une autre approche, **Bremser** [Bremser1978] propose de caractériser des environnements sphériques des atomes et des systèmes cycliques en utilisant un code de sous-structures appelé HORSE. La méthode LaSSI de **Hull** [Hull2001] utilise la valeur de « décomposition singulière » d'un descripteur chimique ou d'une matrice moléculaire en sous-structures pour créer une représentation en moins de dimensions que l'espace chimique original. Ceci permet de calculer la similarité entre deux descripteurs ou entre un descripteur et une molécule.

Xiao [Xiao1997] propose un algorithme qui exploite l'information moléculaire environnant un atome. Ceci se fait couche par couche à partir de l'atome central de la molécule cible, et permet de construire un code structural. Même si l'idée ressemble beaucoup à celle proposée par Dubois, l'algorithme présente des différences significatives dans la manière de coder les fragments obtenus. Ce codage se fait de manière automatique sans prédéfinir à l'avance des fragments spécifiques.

Bender [Bender2004] propose une technique pour la recherche de similarité entre molécules. Les descripteurs utilisés s'appellent des « environnements atomiques » [Xing2002]. Ces descripteurs sont d'interprétation facile et sont très similaires aux « descripteurs de signature moléculaire » [Faulon2003, Faulon2003a]. Ils sont calculés à partir de la table de connectivité. On donne les distances à partir de l'atome <0> et on calcule des vecteurs jusqu'à la distance désirée (dans la figure I.8, jusqu'à une ou deux liaisons). Des *fingerprints* d'environnements moléculaires sont ainsi construits. Ceux-ci sont binaires, pour indiquer la présence/absence de vecteurs de comptage ou de types d'atomes. Cette technique a été utilisée pour retrouver cinq groupes de molécules actives extraits de la base de molécules MDL Drug Data Report (MDDR). Dans une analyse comparative, les auteurs affirment améliorer les résultats obtenus avec des descripteurs 2D et 3D.

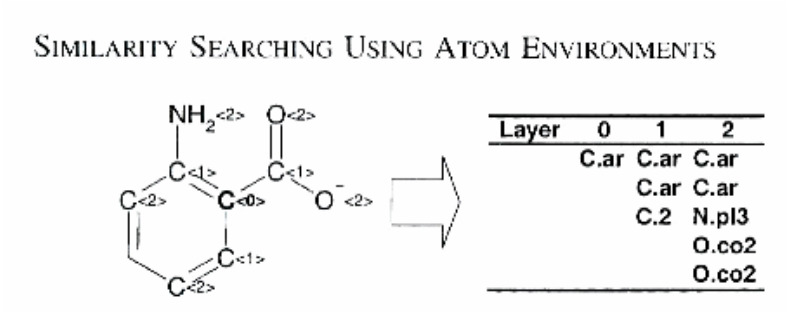


Figure. I.8. Illustration de la génération d'un descripteur atour d'un atome de carbone aromatique.

D'autres contributions qui ont utilisé des descripteurs structuraux de type graphe pour la recherche des molécules ou l'analyse de similarité sont citées dans la littérature [Takahashi1992, Gillet1991, Garey1978]. Une revue des méthodes de recherche qui utilisent des sous-structures a été publiée par **Barnard** [Barnard1993]. Dans ce travail, les avancées quant à l'utilisation des descripteurs structuraux pour la détermination de la similarité et la diversité moléculaires ont été résumées.

I.2.3 La comparaison de descripteurs dans la littérature

Des représentations différentes, outre les descripteurs 2D, ont été le sujet d'études comparatives [Horvath2003, Horvath2003a]. Beaucoup de ces descripteurs ne sont pas très efficaces pour l'analyse de banques de molécules (descripteurs de corrélation, logP, HOMO-LUMO, etc.). D'autres sont adaptés à cet usage sous certaines contraintes de masse, taille ou composition des molécules. Certains sont directement calculables sur la molécule, et d'autres le sont dans un autre espace (WHIM, RDF, etc.). Un échantillon des travaux abordant la comparaison des descripteurs dans un cadre structural est présenté par la suite.

Martin [Martin1998] a comparé la pertinence de différents descripteurs moléculaires. Ils ont trouvé que des descripteurs sous-structuraux simples du type MACCS sont plus puissants pour distinguer les composants actifs des inactifs, par rapport aux *fingerprints* de Daylight. Ils ont également confirmé les relations existantes entre les descripteurs structuraux et les propriétés physicochimiques.

Avec le logiciel **DISSIM** [Flower1998], des études comparatives pour choisir les groupes de descripteurs les plus performants et les moins inter-corrélés ont été effectués. Les résultats incluent des arbres de relations pour 159 descripteurs, pour résoudre le problème de corrélation ainsi que des schémas de poids et de normalisation.

Consoni [Consonni2002a, Consonni2002b] a fait une étude comparative en utilisant trois types différent de descripteurs : descripteurs GETAWAY, descripteurs topologiques du type matrice de Wiener et descripteurs WHIM. Le travail conclut que les descripteurs GETAWAY sont avantageux car ils encryptent l'information 3D, sont facilement calculables et permettent de bonnes prédictions de propriétés physicochimiques.

Feng [Feng2003] a comparé différents types de descripteurs (1D, 2D et 3D) en utilisant quatre types de bases de molécules différentes et trois méthodes statistiques. Il a conclu qu'il n'y avait pas de différences de performance significatives entre ces descripteurs.

Hicks [Hicks1990] a évalué la performance et l'efficacité de cinq systèmes de recherche basés sur les sous structures: MACCS, DARC, HTSS, CAS Registry MVSSS et S4. Les résultats ont montré que tous les systèmes donnent des résultats similaires en termes de performance, sauf S4 qui présente des temps de calcul plus longs.

Martin [Martin2001] a effectué une étude pour sélectionner les descripteurs moléculaires les plus pertinents pour des tests biologiques. Ils ont utilisé la méthode de Ward [Brown1996] pour regrouper les molécules actives et testé trois méthodes de codage chimique 2D et trois de codage 3D. Ses résultats indiquent que les descripteurs structuraux 2D et 3D peuvent contenir de l'information recoupée. Mais des molécules qui semblent être similaires en 2D, peuvent être différentes en 3D si l'on considère leurs propriétés liées aux récepteurs biologiques. L'incrémentation de la diversité dans une base de test devrait donc augmenter les chances de trouver de nouvelles molécules intéressantes.

Les travaux rapportés par **Good** [Good1998] résument une série de techniques utiles pour quantifier explicitement la similarité moléculaire en 3D. Les calculs ont été faits en utilisant des descripteurs

de forme moléculaire et des MEP. De nombreuses propriétés moléculaires, indices et protocoles ont été ainsi présentés et discutés.

Godden [Godden2000] propose une méthode pour calculer et comparer la variabilité des descripteurs moléculaires utilisés en bases de données moléculaires. Son analyse est basée sur des histogrammes qui contiennent la distribution de descripteurs moléculaires et le calcul de l'entropie de Shannon (laquelle reflète la variabilité du descripteur). Des différences significatives ont été observées et l'entropie de Shannon s'est révélée être un facteur discriminant efficace.

Il est important de noter que plusieurs travaux [Martin2001, Barnard1993 et Bayada1999] affirment que les descripteurs sous-structuraux ont de meilleurs rendements dans le criblage de bases de données moléculaires et permettent souvent d'établir des relations entre les molécules et des propriétés biologiques données. La question de savoir pourquoi ces descripteurs ont une meilleure performance a été abordée par **Martin** [Martin2001]. Dans ces travaux, des propriétés physiques calculées ont été utilisées, au lieu des activités biologiques usuelles. Des exercices de regroupement de molécules pour tester la performance des descripteurs ont permis de démontrer que les descripteurs sous-structuraux contiennent des informations sur les propriétés physicochimiques et des caractéristiques 3D dans une proportion équilibrée qui permet la prévision des activités biologiques [White2003]. Une des conclusions des travaux de **Bayada** [Bayada1999] concerne les descripteurs sous-structuraux. Il démontre que ces descripteurs sont très performants et établissent des relations entre les molécules et des propriétés biologiques données. Dans ce travail, environ la moitié des descripteurs initialement considérés a été éliminée plus tard. Beaucoup de descripteurs traditionnellement utilisés pour des études QSAR ont été inefficaces pour des analyses de diversité. Seule l'utilisation des *fingerprints* et de descripteurs englobant la molécule entière a donné des résultats supérieurs à la sélection aléatoire dans un groupe de diverses drogues potentielles.

D'autre part, **Makara** [Makara2001] affirme que les méthodes 2D, en comparaison avec les méthodes 3D, souffrent de beaucoup d'inconvénients. Entre autres, sont énumérés : le manque

d'information sur la forme de la molécule, la localisation des groupes fonctionnels dans l'espace, la mauvaise reconnaissance d'isomères et l'absence de traitement de problèmes conformationnels.

Une solution alternative est proposée par **Schuffenhauer** [Schuffenhauer2000] qui suggère une combinaison de descripteurs 2D et 3D. Ses résultats montrent que ceux-ci ont une meilleure performance par rapport au groupe de descripteurs 2D pour retrouver des molécules dans une base de données BIOSTER. Une autre possibilité est présentée par **Sun** [Sun2004] qui propose un descripteur moléculaire universel pour prédire des propriétés ADME. Il affirme que les descripteurs 1D, 2D et 3D ont des difficultés pour codifier les informations pertinentes de la molécule. Il propose l'utilisation de variantes du *fingerprint,* en utilisant la classification des type d'atomes comme un moyen de description insuffisamment exploité jusqu'à aujourd'hui selon l'auteur.

Conclusion

Dans ce chapitre, nous avons introduit les concepts et l'historique de quelques techniques de criblage virtuel et de criblage de haut débit. L'interrogation des bases de données et l'analyse de la similarité et de la diversité des molécules ont été au centre de notre analyse. Différents travaux de comparaison des descripteurs ont été discutés, notamment ceux en rapport avec les descripteurs structuraux. Leurs conclusions nous mènent à considérer l'approche des sous-structures comme une voie viable pour décrire les molécules dans le cadre de criblage virtuel qui relève de notre problématique.

Références

[Bajorath2002] Bajorath, J., *Integration of Virtual and High-Throughput Screening*. Nature Reviews, 1 (2002) 882-894.

[BajorathWeb] Bajorath, J., *Virtual Screening in drug discovery: Methods, expectations and reality*. Information disponible à : http://www.currentdrugdiscovery.com

[Bayada1999] Bayada, D.M., Hamersma, H., Van Geerestein, V.J., *Molecular Diversity and Representativity in Chemical Databases*, J. Chem. Inf. Comput. Sci., 39 (1999) 1-10.

[Barnard1993] Barnard, J.M., *Substructure Searching Methods: Old and New*, J. Chem. Inf. Comput. Sci., 33 (1993) 532-538.

[Bender2004] Bender, A., Mussa, H.Y., Glen, R.C., *Molecular Similarity searching using atoms environments, information-based feature selection and a naïve Bayesian classifier*, J. Chem. Inf. Comput. Sci. 44 (2004) 170-178.

[Bocker2004] Böcker, A., Schneider, G., Teckentrup, A., *Status of HTS Data mining approaches*, QSAR Comb. Sci. 23 (2004) 207-213.

[Bremser1978] Bremser, W., *Horse- A novel substructure code*, Anal. Chem. Acta., 103 (1978) 355-365.

[Brown1996] Brown, R.D., Martin, Y.C., *Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compounds selection*, J. Chem. Inf. Comput. Sci., 36 (1996) 572-584.

[Cuissart2002] Cuissart, B., Touffet, F., Crémilleux, B., Bureau, R., Rault, S., *The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/ biodegradability relationships*, J. Chem. Inf. Comput. Sci., 42 (2002) 1043-1052.

[Consonni2002a] Consonni, V., Todeschini, R., Pavan, M., *Structure/Response correlation and Similarity/Diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors*, J. Chem. Inf. Comput. Sci., 42 (2002) 682-692.

[Consonni2002b] Consonni, V., Todeschini, R., Pavan, M., *Structure/Response correlation and Similarity/Diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies*, J. Chem. Inf. Comput. Sci., 42 (2002) 693-705.

[Doucet1998] Doucet, J.P., Panaye, A., *3D Structural Information: form property prediction to substructure recognition with neural networks*, SAR and QSAR Envirom. Res., 8 (1998) 249-272.

[Dubois1986] Dubois, J.E., Mercier, C., Panaye, A., *DARC topological system and computer aided design*, Acta Pharm. Jugosl., 36 (1986) 135-169.

[Dubois1999] Dubois, J.E., Doucet, J.P., Panaye, A., Fan, B.T., *DARC site toplogical correlations: ordered structural descriptors and property evaluation.* In Devillers, J. and Balaban, T. (Eds). Topological indices and related descriptors in QSAR and QSPR, Gordon and Breach Sciences Publishers, Amsterdam, 1999, pp. 613-673.

[Faulon2003] Faulon, J.L., Visco, D.P. Jr, Pophale, R.S., *The signature Molecular Descriptor. 1. Using extended valence sequences in QSAR and QSPR studies*, J. Chem. Inf. Comput. Sci., 43 (2003) 707-720.

[Faulon2003a] Faulon, J.L., Churchwell, C.J., Visco, D.P Jr., *The signature Molecular Descriptor. 2. Enumerating molecules from their extended valence sequences,* J. Chem. Inf. Comput. Sci., 43 (2003) 721-734.

[Flower1998] Flower, D.R., *DISSIM: a program for the analysis of chemical diversity*, J. Molec. Graph. Mod., 16 (1998) 239-253.

[Feng2003] Feng, J., Lurati, L., Ouyang, H., *Predictive toxicology : benchmarking molecular descriptors and statistical methods.* J. Chem. Inf. Comput. Sci. 43 (2003) 1463-1470.

[Garey1978] Garey, M.G., Johnson, D.S., *Computers and Intractability, a Guide to the Theory of NP-Completeness*, In Klee V. (Ed.) A series of books in the Mathematical Sciences, W.H. Freeman and company, New York, 1978, pp. 202-205.

[Gillet1991] Gillet, V.J., Downs, G.M., Holliday, J.D., Lynch, M.F., Dethlefsen, W., *Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph generation*, J. Chem. Inf. Comput. Sci., 31 (1991) 260-270.

[Gillet2003] Gillet, V., Willett, P., Bradshaw, J., *Similarity Searching Using Reduced Graphs*, J. Chem. Inf. Comput. Sci., 43 (2003) 338-345.

[Good1998] Good, A.C., Richards, W.G., *Explicit calculation of 3D molecular Similarity*, Perspectiv. Drug Disc. Design, 9/10/11 (1998) 321-338.

[Godden2000] Godden, J.W., Stahura, F.L., Bajorath, J., *Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations.* J. Chem. Inf. Comput. Sci., 40 (2000) 796-800.

[Gordon1998] Gordon E. M., Kerwin, J.F. Jr (Eds.) Combinatorial Chemistry and Molecular Diversity in Drug Discovery, Wiley & Sons, New York, 1998.

[Hicks1990] Hicks, M.G., Jochum, C., *Substructure search systems. 1. Performance comparison of the MACCS, DARC, HTSS, CAS Registry MVSSS and S4 Substructure search systems,* J. Chem. Inf. Comput. Sci., 30 (1990) 191-199.

[Horvarth2003] Horvarth, D., Jeandenans, C., *Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces - A novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles.* J. Chem. Inf. Comp. Sci., 43 (2003) 680-690.

[Horvarth2003a] Horvath, D., Jeandenans, C., *Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces - A Benchmark for neighborhood behavior assessment of different in silico similarity metrics.* J. Chem. Inf. Comp. Sci,, 43 (2003) 691-698.

[Hull2001] Hull, R.D., Singh, S.B., Nachbar, R.B., Sheridan, R.P., Kearsley, S.K., Fluder, E.M., *Latent Semantic Structure Indexing (LaSSI) for defining chemical similarity*, J. Med. Chem., 44 (2001) 1177-1184.

[Ivanciuc2000] Ivanciuc, O., Taraviras, S.L., Cabrol-Bass, D., *Quasi-orthogonal basic sets of molecular graphs descriptors as a chemical diversity measure*, J. Chem. Inf. Comput. Sci., 40 (2000) 126-134.

[Japertas2002] Japertas, P., Didziapetris, R., Petrauskas, A., *Fragmental Methods in the design of new compounds. Applications of the Advanced Algorithm Builder*, QSAR, 21 (2002) 23-37.

[Johnson1990] Johnson, A.M., Maggiora, G.M. (Eds.) Concepts and Applications of Molecular Similarity, John Willey & Sons, New York, Inc. 1990.

[Kopp1842] Kopp, H., Ann. Chem. 41 (1842) 79. Reedited in 1954 as, Kopp, H. Ann. Annalen der Chemie und pharm, 92 (1854) 1.

[Lengauer2004] Lengauer, T., Lemmen, C., Rarey, M., Zimmermann, M. *Novel Technologies for Virtual Screening.* Drug Disc. Today, 1 (2004) 27-33.

[Martin1998] Martin, Y.C., Bures, M.G., Brown, R.D., *Validated Descriptors for Diversity Measurements and Optimization*, Pharm. Pharmacol. Commun., 4 (1998) 147-152.

[Martin2001] Martin Y. C., *Molecular Diversity: how we measure it? Has it lived up to its promise?*, Il Farmaco 56 (2001) 137-139.

[Martin2002] Martin, Y.C., Kofron, J.L., Traphagen, L.M. *Do structurally similar molecules have similar biological activity?,* J. Med. Chem., 45 (2002) 4350-4358.

[Makara2001] Makara G., *Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity*, J. Med. Chem., 44 (2001) 3563-3571.

[Moos1996] Moos W.H., *Combinatorial Chemistry: a "Molecular Diversity Space" Odyssey Approaches 2001*, Pharmaceutical News, 3 (1996) 23-26**.**

[Nikolova2003] Nikolova, N., Jaworska, J., *Approaches to Measure Chemical Similarity - a Review,* QSAR Comb. Sci., 22 (2003) 1006-1026.

[OFarrell2005] O'Farrell, M., Lewis, E., Flanagan, C., Lyons, W., Jackman, N., *Comparison of k-NN and neural network methods in the classification of spectral data from an optical fibre-based sensor system used for quality control in the food industry*. Sensors and Actuators B: Chemical, 111-112 (2005) 354-362.

[Randic1979] Randic, M., Wilkins, C.L., *Graph theoretical ordering of structures as a basis for systematic searches for regularities in molecular data,* J. Phys. Chem., 83 (1979) 1525-1540.

[Randic2001] Randic, M., *Graph valence shells as molecular descriptors,* J. Chem. Inf. Comput. Sci., 41 (2001) 627-630.

[Richardson1876] Richardson B.W., The diseases of modern life, London, Macmillan, 1876.

[Schuffenhauer2000] Schuffenhauer, A., Gillet, V.J., Willett, P., *Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors,* J. Chem. Inf. Comput. Sci., 40 (2000) 295-307.

[Sun2004] Sun, H., *A universal molecular descriptor system for prediction of logP, logS, logBB and absorption*, J. Chem. Inf. Comput. Sci., 44 (2004) 748-757.

[Stahura2004] Stahura, F.L., Bajorath, J. *Virtual screening methods that complements HTS*. Comb. Chem. & HTS, 7 (2004) 259-269.

[Stu2003] Stu Borman, *The many faces of combinatorial chemistry*, Chem. Engin. News, 81 (2003) 45-56.

[Takahashi1992] Takahashi, Y., Sukekawa, M., Sasaki, S., *Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical Structur*e, J. Chem. Inf. Comput. Sci., 32 (1992) 639-643.

[Walters1998] Walters, W.P., Stahl, M.T., Murcko, M.A. *Virtual Screening - An Overview*, Drug Discovery Today, 3 (1998) 160-178.

[White2003] White, M., Willett, P., *Evaluation of Similarity Measures for Searching the Dictionary of* Natural Products *Database*, J. Chem. Inf. Comput. Sci., 43 (2003) 449-457.

[Whitley2000] Whitley, D.C., Ford, M.G., Livingstone, D.J., *Unsupervised forward selection: a method for eliminating redundant variables*, J. Chem. Inf. Comput. Sci., 40 (2000) 1160-1168.

[Willett1997] Willett, P., *Using Computational Tools to Analyze Molecular Diversity,* In DeWitt, H., Czarnik, A.W. (Eds.) Combinatorial Chemistry; A Short Course, American Chemical Society Books, Washington DC, 1997.

[Weber2000] Weber, L., *High-diversity combinatorial libraries*, Curr. Op. Chem. Bio., 4 (2000) 295-302.

[Xiao1997] Xiao, Y., Qiao, Y., Zhang, J., Lin, S., Zhang, W., *A method for substructure search by atom-centered multilayer code*, J. Chem. Inf. Comput. Sci., 37 (1997) 701-704.

[Xing2002] Xing, L.,Glen, R.C., *Novel methods for the prediction of Log P, pKa and Log D*, J. Chem. Inf. Comput. Sci., 42 (2002) 796-805.

CHAPITRE II.

BASES DE DONNEES:

REPRESENTATION ET

STRUCTURATION

II.1. Bases de données. Lexique et construction
II.1.1 Lexique et format de molécules
II.1.2 Construction de la base de données de fragments (FragDB)
II.1.2.1 Les atomes génériques
II.1.2.2 L'origine des fragments et des sous-structures
II.1.2.3 Un aperçu des bases de fragments
II.1.3 Construction des bases de données QueryDB et TestDB
II.2. Structuration des informations moléculaires et XML
II.2.1 Les langages de marquage
II.2.1.1 Histoire
II.2.1.2 Principes
II.2.1.3 XML pour structurer les informations chimiques
II.2.2 La structuration de la FragDB avec XML
II.2.2.1 Création et remplissage d'un index-XML de fragments
II.2.2.2 Une DTD pour valider l'index-XML
II.2.3 La structuration du QueryDB et du TestDB
II.2.3.1 Transformation des molécules et création du VecteurRepresentatif-XML
II.2.3.2 Une DTD pour valider le VecteurRepresentatif-XML
II.2.3.3 Une DTD pour valider l'indexResult-XML
II.2.4 La représentation des connaissances

CHAPITRE II. BASES DE DONNEES : REPRESENTATION ET STRUCTURATION

Les sections I.1 et I.2 nous ont permis de faire le tour des approches utilisant des descripteurs structuraux 2D et des raisons de les adopter. Entre autres avantages, ont été nommées leur capacité à coder des propriétés physicochimiques, leur facilité d'utilisation et d'implémentation, la diversité des niveaux de complexité disponibles ainsi que la présence des informations 3D implicites dans les modèles. Ce sont les mêmes raisons qui nous ont amené à adopter des descripteurs sous-structuraux pour la construction et la structuration d'un ensemble de bases de données chimiques rassemblant les informations nécessaires à notre outil de criblage virtuel et d'analyse de similarité moléculaires.

II.1. Bases de données. Lexique et construction

Une base de données regroupe un ensemble d'informations organisées de manière à faciliter l'exploitation des connaissances inhérentes aux éléments qui la composent. La base doit avoir le minimum de redondance dans une taille maximale. Elle doit permettre le partage des informations et garantir l'intégrité des données. En informatique le modèle de base de données prédominant est le modèle relationnel (et ses multiples variantes). Dans une base de données relationnelle les données sont organisées en forme de tables. Chaque table contient des champs typés (des champs dont on connaît le type d'information contenue). Pour effectuer des requêtes on peut faire la jonction des tables (caractéristique novatrice des bases de données relationnelles par rapport aux systèmes de fichiers) et utiliser des filtres sur l'information souhaitée.

D'une manière générale en chimie, les données peuvent être de nature très différente. Celles-ci comprennent : des propriétés physicochimiques (nombres entiers ou réels, valeurs binaires), des variations sur la forme ou l'apparence (graphes, table de connectivité, 2D, 3D, etc.), des propriétés

électroniques (conformations, énergies, etc.), des données spectroscopiques (IR, Raman, UV), etc. La diversité des informations moléculaires a donné lieu à une grande variété de représentations chimiques par ordinateur. Dans notre cas, la représentation d'une molécule se fera par rapport aux fragments la constituant et à leurs propriétés implicites. Nous montrerons ensuite la manière dont nous avons construit et structuré nos bases de données moléculaires.

II.1.1 Lexique et format de molécules

Tout au long de ce manuscrit, une série de termes et d'abréviations propres à notre logiciel ainsi qu'une nomenclature particulière pour la base de fragments sera introduite. Comme pour tout logiciel de criblage virtuel de haut débit, notre logiciel utilise plusieurs bases de données, structurées à différents niveaux et avec des buts différents.

La base de molécules composée des molécules cibles est appelé « QueryDB » et la base de molécules à comparer : « TestDB ». Une fois que l'utilisateur à choisi le deux bases « QueryDB », « TestDB », l'analyse de similarité sera effectuée en utilisant une base de sous-structures prédéfinies manuellement et qui sera nommée « FragDB ». La figure II.1 montre la composition des bases de données de l'outil.

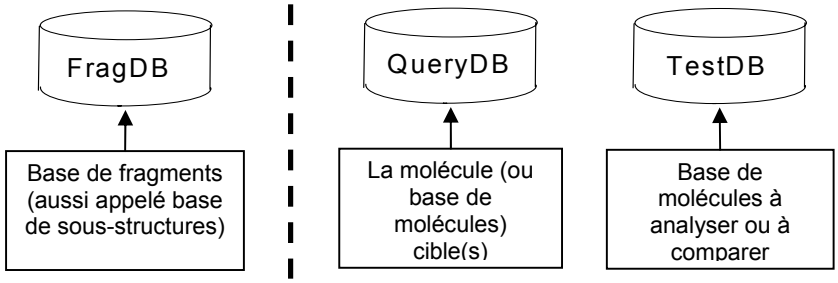

FragDB
QueryDB
TestDB
Base de fragments (aussi appelé base de sous-structures)
La molécule (ou base de molécules) cible(s)
Base de molécules à analyser ou à comparer

Figure II.1. Lexique utilisé pour désigner les bases des molécules utilisées dans l'outil de criblage.

Plus tard, dans la section de structuration, nous travaillerons avec des fichiers de structuration de données. Ces fichiers (par exemple « index.xml », « indexResult.xml », etc.) codent l'information chimique des fragments ou des molécules. De la même manière, les noms de fichiers des sous-structures composant la FragDB seront désignés avec un nom spécifique codant des informations chimiques. Ceci sera expliqué dans le chapitre suivant.

Pour l'acquisition des données chimiques, les molécules et les fragments devront être en format .MOL. Un fichier en format .MOL peut mémoriser des informations sur les atomes et les liaisons d'une molécule en 2D ou en 3D, ainsi que les caractéristiques d'une réaction chimique. Après un bloc d'en-tête du fichier, le contenu principal du fichier .MOL consiste en informations sur la connectivité, et sur la nature des atomes et des liaisons. Ce format sera présenté en détail dans l'annexe 3.

Il est important de noter que pendant les trente dernières années le traitement des informations chimiques à donné lieu à de très nombreux formats de représentation de molécules. Du fait que ces différents travaux ont été conduits sans prédéfinir une norme standard, plusieurs formats co-existent. Parmi eux, les formats plus populaires sont : SMD [Bebak1989] (qui est recommandé par la CAS), MOL [Dalby1992] proposé par MDL, SMILES [Weininger1988, Weininger1989], WLN [Wiswesser1954], DARC [Dubois1986, Dubois1999], etc.

II.1.2 Construction de la base de données de fragments (FragDB)

Comme indiqué dans la section précédente, la FragDB consiste en une base de sous-structures moléculaires prédéfinies manuellement. Les définitions structurales et les critères de construction de la base seront présentés ci-dessous. Pour prendre en compte la plus grande diversité chimique dans les fragments à définir (sans toutefois être exhaustif), nous avons utilisé des atomes génériques dans la construction des sous-structures.

II.1.2.1 Les atomes génériques

Ces atomes génériques respectent une hiérarchie définie à partir des modèles déjà existants et des besoins particuliers de notre outil. Ainsi, une premier classe d'atomes appelé « * » représente tous les atomes de la table périodique moderne à nos jours. Même si cet atome générique n'est pas proprement inclus dans les sous-structures il permet de définir un cadre pour classer tous les autres atomes. Au deuxième niveau de complexité nous avons trois classes d'atomes : les carbones aromatiques « A », l'hydrogène « H » (non inclus de manière explicite dans les représentations des sous-structures) et tout les autres atomes représentés par « Q ». La classe « Q » elle-même est composée des halogènes « X », des métaux « M » et des hétéroatomes importants en chimie pharmaceutique « Z », à savoir les atomes de bore, d'oxygène, d'azote, de phosphore et de soufre. Le reste des éléments est inclus dans la classe « R », voir la figure II.2 et le tableau II.1.

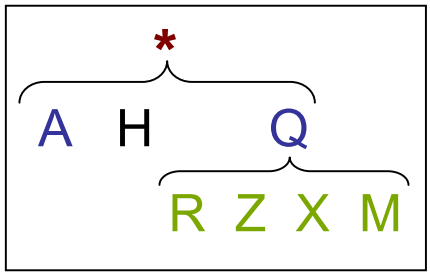


Figure. II.2. Hiérarchie proposée des atomes génériques pour la structuration de la base de fragments. Le niveau le plus général est « * ». Les carbones aromatiques « A » et l'hydrogène « H » sont mis à part. Pour finir le groupe « Q » est décomposé en atomes métalliques « M », atomes non métalliques importants « Z », halogènes « X » et le reste des atomes « R » (dont l'atome de C non aromatique). De façon générale l'atome d'Hydrogène n'est pas explicite.

Le tableau II.1 montre en détail les atomes particuliers inclus dans chaque catégorie d'atomes génériques. Les éléments pris en compte appartiennent à la table périodique actuelle [PerTableWeb].

Symbole	Atomes Représentés
*	Tous les éléments de la table périodique moderne = "H","He","Li","Be","B","C","N","O","F","Ne", "Na","Mg","Al","Si","P","S","Cl","Ar","K","Ca", "Sc","Ti","V","Cr","Mn","Fe","Co","Ni","Cu","Zn", "Ga","Ge","As","Se","Br","Kr","Rb","Sr","Y","Zr", "Nb","Mo","Tc","Ru","Rh","Pd","Ag","Cd","In","Sn", "Sb","Te","I","Xe","Cs","Ba","La","Ce","Pr","Nd", "Pm","Sm","Eu","Gd","Tb","Dy","Ho","Er","Tm","Yb", "Lu","Hf","Ta","W","Re","Os","Ir","Pt","Au","Hg", "Tl","Pb","Bi","Po","At","Rn","Fr","Ra","Ac","Th", "Pa","U","Np","Pu","Am","Cm","Bk","Cf","Es","Fm", "Md","No","Lr","Rf","Db","Sg","Bh","Hs","Mt","Ds", "Rg","Uub","Uut","Uuq","Uup","Uuh","Uus","Uuo"
A	Atomes aromatiques
Q	Tous les éléments à l'exception de H et de A = "He","Li","Be","B","C","N","O","F","Ne", "Na","Mg","Al","Si","P","S","Cl","Ar","K","Ca", "Sc","Ti","V","Cr","Mn","Fe","Co","Ni","Cu","Zn", "Ga","Ge","As","Se","Br","Kr","Rb","Sr","Y","Zr", "Nb","Mo","Tc","Ru","Rh","Pd","Ag","Cd","In","Sn", "Sb","Te","I","Xe","Cs","Ba","La","Ce","Pr","Nd", "Pm","Sm","Eu","Gd","Tb","Dy","Ho","Er","Tm","Yb", "Lu","Hf","Ta","W","Re","Os","Ir","Pt","Au","Hg", "Tl","Pb","Bi","Po","At","Rn","Fr","Ra","Ac","Th", "Pa","U","Np","Pu","Am","Cm","Bk","Cf","Es","Fm", "Md","No","Lr","Rf","Db","Sg","Bh","Hs","Mt","Ds", "Rg","Uub","Uut","Uuq","Uup","Uuh","Uus","Uuo"
M	Atomes métalliques = "Al","Sc","Ti","V","Cr","Mn","Fe","Co","Ni","Cu", "Zn","Ga","Y","Zr","Nb","Mo","Tc","Ru","Rh","Pd", "Ag","Cd","In","Sn","Lu","Hf","Ta","W","Re","Os", "Ir","Pt","Au","Hg","Tl","Pb","Bi","Po","Lr","Rf", "Db","Sg","Bh","Hs","Mt","Ds","Rg","Uub"
X	Halogènes = "F","Cl","Br","I"
Z	Atomes non métalliques importants = "B","N","O","P","S"
R	Tous les autres atomes = "He","Li","Be","C","Ne","Na","Mg","Si","Ar","K","Ca", "Ge","As","Se","Kr","Rb","Sr","Sb","Te","Xe","Cs","Ba", "La","Ce","Pr","Nd","Pm","Sm","Eu","Gd","Tb","Dy", "Ho","Er","Tm","Yb","At","Rn","Fr","Ra","Ac","Th", "Pa","U","Np","Pu","Am","Cm","Bk","Cf","Es","Fm", "Md","No,"Uut","Uuq","Uup","Uuh","Uus","Uuo"

Tableau II.1. Détail des atomes inclus dans les catégories d'atomes génériques de la figure II.2.

La catégorie d'atomes métalliques (M) a été construite en prenant en compte les éléments à fort caractère métallique. La catégorie des atomes non métalliques (Z) dits « importants » a été définie sur la base de leur fréquence d'apparition reportée dans la littérature (voir les travaux reportés dans le chapitre II.1.2.2 : [Erl2003, Stobaugh1988, Xu2000]). Finalement, on inclut l'atome de Carbone dans la catégorie R (car traditionnellement, les chaînes -R- représentent des chaînes aliphatiques).

Comme tout outil traitant des atomes et de l'information chimique, nous travaillons avec l'information du numéro atomique des éléments. Nous avons donc eu besoin d'assigner des « numéros atomiques » fictifs aux atomes génériques (tableau II.2).

Numéro Atomique	Symbole	Valeur
150	*	Tous les éléments
148	A	Carbone aromatique
146	Q	Tous les éléments excepté H et A
144	M	Eléments métalliques
142	X	Eléments halogènes
140	Z	Eléments non métalliques
138	R	Le reste des éléments
119-137	-	Numéros non assignés
1-118	H - Uuo	Eléments de la table périodique actuelle

Tableau II.2. Eléments et numéros atomiques correspondants.

L'assignation de numéros atomiques fictifs aux atomes génériques a été nécessaire pour leur détection et traitement futur au sein de l'outil. Les éléments de numéro atomique de 1 à 118, gardent leur valeur traditionnelle. Les numéros qui vont de 118 à 137 n'ont pas encore été attribués. Les numéros qui vont de 138 à 150 sont des extensions (numéros atomiques fictifs) assignés aux atomes génériques.

II.1.2.2 L'origine des fragments et des sous-structures

Souvent, les termes « fragment » et « sous-structure » sont utilisés de façon interchangeable dans la

littérature chimique actuelle. Toutefois certaines subtilités font état de différence entre ces deux concepts : une « sous-structure » est définie comme toute partie d'une molécule, composée d'au moins deux atomes et une liaison et qui ne contient *per-se* une connotation quelconque. Un « fragment » est une sous-structure à laquelle on attache un sens utilitaire, une propriété, ou un but structural d'intérêt moléculaire. Dans la suite de ce manuscrit nous utiliserons indifféremment ces termes.

Pour effectuer la construction de la FragDB il faut donc remonter à la conception même de groupe fonctionnel. Un groupe fonctionnel est une sous-structure qui a une connotation d'activité, ou une possible interaction avec un site actif. Ces fragments vont former les éléments constituants d'une molécule et seront en conséquence les éléments constitutifs des vecteurs descripteurs de la dite molécule. C'est pour cette raison qu'il est important de choisir d'une manière optimale la composition de la FragDB. La qualité des descripteurs moléculaires dépendra en grande partie de la composition de cette base car pour calculer des ressemblances entre molécules nous utilisons comme critère principal leurs éléments structuraux.

Comme nous l'avons expliqué dans le chapitre I, le « principe de similarité des propriétés », est à la base même de notre approche. Donc, des molécules structuralement similaires auront plus de chances de se comporter de manière similaire.

Etre « structuralement similaire » signifie en langage chimique, partager les mêmes fragments ou sous-structures et plus important encore, les même groupes fonctionnels. Nous avons donc cherché, sans vouloir être exhaustifs, quels étaient les groupes fonctionnels les plus courants et les plus importants.

Quelques travaux ont fait l'étude de la fréquence d'apparition de sous-structures, fragments, atomes, etc. extraites des bases de données moléculaires et ont publié des listes détaillées et ordonnées de ces données.

Un des travaux de référence est l'article du CAS [Stobaugh1988]. Dans cet article, les statistiques de la base de *CAS Registry System* pour la fréquence d'apparition des substances, des systèmes cycliques et des éléments sont présentées. On remarque déjà, à l'époque, l'abondance des systèmes cycliques

(80%). En plus, on remarque l'augmentation avec le temps des systèmes bicycles et monocycles (32% et 45% respectivement). Sans surprise les éléments les plus répandus sont (dans l'ordre): H, C, O, N, S, Cl, etc. Un travail plus récent de Xu [Xu2000] fait l'analyse de cinq bases de données courantes en chimie médicinale, pour construire un index qui déterminera le degré de ressemblance d'une molécule à une drogue. Même si le *Top-10* des systèmes cycliques coïncide avec celui de l'étude CAS, on observe que les éléments les plus utilisés sont : H, C, O, Cl, N, F, etc. Dans une autre étude [Ertl2003] des substituant organiques sont extraits à partir d'une base de plus de 3 millions de molécules fournis par Novartis. Mise à part les applications particulières de ces résultats (construction d'un outil de bioisosterisme, calcul de la diversité moléculaire), nous les avons utilisés pour aider à la conception de notre liste de fragments représentatifs.

La construction de la FragDB s'est donc effectuée en plusieurs étapes. Dans un premier temps, on a consulté les références listées plus haut de manière à inclure des sous-structures courantes et fréquentes dans la base. Ensuite, des sous-structures intéressantes issues de la bibliographie ont été relevées à la main et on a complété la liste principale avec des sous-structures d'intérêt pharmaceutique et médical. Toute cette information a été confrontée à l'expertise d'un chimiste pharmaceutique.

La base comptait alors près de 500 fragments, qui ont été ensuite rassemblés dans le but d'établir différents niveaux de granularité au moment de retrouver les fragments. L'importance de l'existence des niveaux de granularité sera abordée dans le chapitre III lors de l'explication du processus de comparaison de structures moléculaires et de reconnaissance des motifs structuraux.

Il est important de noter que des outils d'extraction automatique de fragments ont été présentés dans la communauté [Dubois1980a, Dubois1980b, Carabedian1988, Bremser1978], mais on observe souvent qu'un grand nombre des sous-structures générées sont recouvrantes.

Une fois que les sous-structures ont été choisies et définies en utilisant les atomes génériques décrits plus haut, on a nommé et indexé les fichiers dans la FragDB. Pour cela on a trouvé adéquat de nommer

les fragments en suivant une « nomenclature » (voir figure II.3) qui code des informations chimiques difficiles à structurer plus tard, comme les concepts d'aromaticité ou de mélange d'hétéroatomes.

Les informations codées sous le nom de fichier des fragments nous permettront de compléter notre structure de données chimiques et d'améliorer la recherche de fragments et le criblage des molécules.

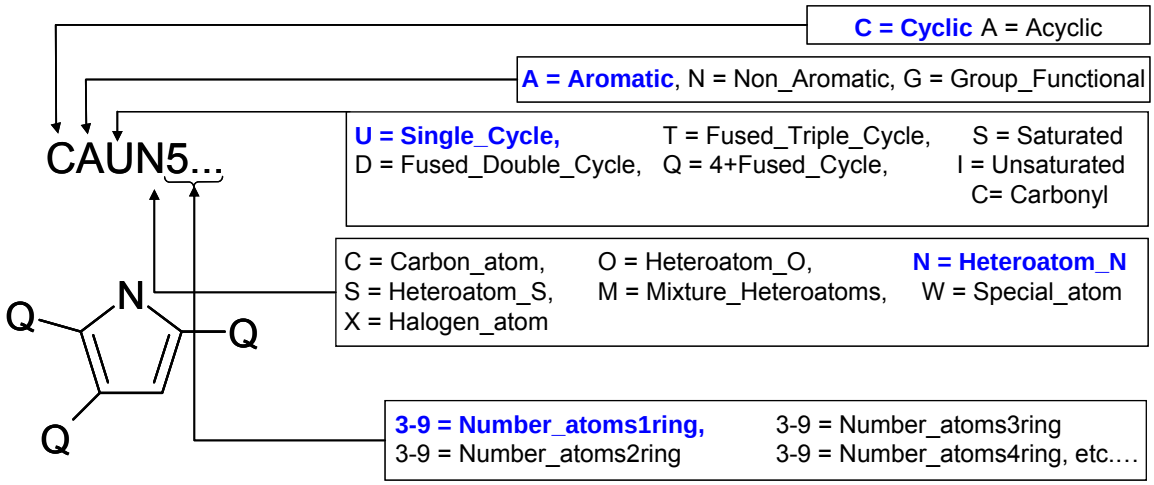

C = Cyclic A = Acyclic
A = Aromatic, N = Non_Aromatic, G = Group_Functional
CAUN5...
U = Single_Cycle, T = Fused_Triple_Cycle, S = Saturated
D = Fused_Double_Cycle, Q = 4+Fused_Cycle, I = Unsaturated
C= Carbonyl
C = Carbon_atom, O = Heteroatom_O, N = Heteroatom_N
S = Heteroatom_S, M = Mixture_Heteroatoms, W = Special_atom
X = Halogen_atom
N
Q
Q
Q
3-9 = Number_atoms1ring, 3-9 = Number_atoms3ring
3-9 = Number_atoms2ring 3-9 = Number_atoms4ring, etc.…

Figure II.3. Exemple de la signification des lettres et des chiffres composants le nom de fichier des éléments de la FragDB.

Comme montre la figure II.3, deux grands groupes de fragments constituent FragDB, les fragments cycliques (C) et les fragments acycliques (A).

Dans la catégorie Acyclique, nous avons regroupé principalement les fragments par famille de groupes fonctionnels (AG), la catégorie AN restant toutefois possible, particulièrement pour décrire les fragments acycliques simples du type C=C, Cl - N, et qui n'appartiennent pas à une catégorie de groupe fonctionnel.

La catégorie C est divisée en cycles aromatiques (CA) et non aromatiques (CN). Dans les cycles avec des hétéroatomes, les lettres S, O, N, M se chargent de designer leur apparition. La plupart des cycles

liés et pontés sont analysés et séparés préalablement à la comparaison avec FragDB, en conséquence les catégories D, T et Q sont restés hors usage après l'implémentation de l'outil de coupure.

A la fin du nom de fichier on observe également, un code alphanumérique qui désigne un nombre d'usage interne. Ce code permet de regrouper les molécules par famille. Ainsi pour une sous-structure avec nom de fichier : CAUN5-156Qb, « CAUN5 » désigne un cycle aromatique à cinq nœuds avec un azote, le code « 156 » désigne la famille des pyrroles, et « Qb » nous indique que il est substitué dans 3 nœuds (voir figure II.3).

Nous avons pris la précaution de limiter les combinaisons possibles, pour éviter des contradictions chimiques ou de non-sens. Par exemple, un fragment ne pourra jamais s'appeler «AA… » car la condition pour qu'une molécule soit aromatique est qu'elle soit cyclique. Les combinaisons possibles de noms de fichiers sont réduites à celles montrées dans les figures suivantes.

II.1.2.3 Un aperçu des bases de fragments

Dans notre outil de criblage virtuel, quatre bases de données FragDB ont été construites et ordonnées selon les critères de la section II.1.2.2. Ces bases sont associées aux différents types d'informations structurales à traiter. Leur classification obéie à la nomenclature montrée dans les figures II.4 et II.5.

La FragDB concerne actuellement :

1. 60 fragments contenant des cycles aromatiques (CA),
2. 450 fragments contenant des cycles non aromatiques (CN),
3. 11 fragments contenant des chaînes acycliques (AN),
4. 50 fragments contenant des groups fonctionnels (AG).

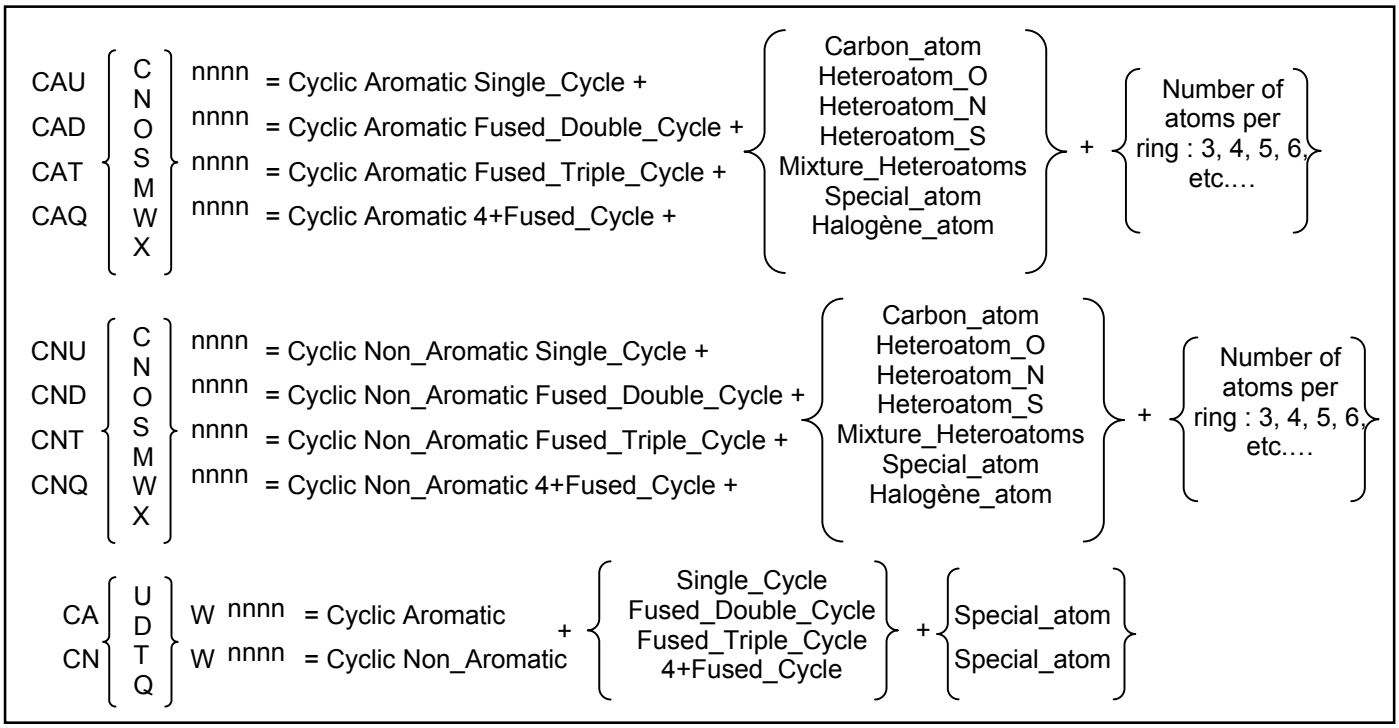


Figure II.4. Combinaisons possibles de noms de fichier de fragments cycliques dans FragDB

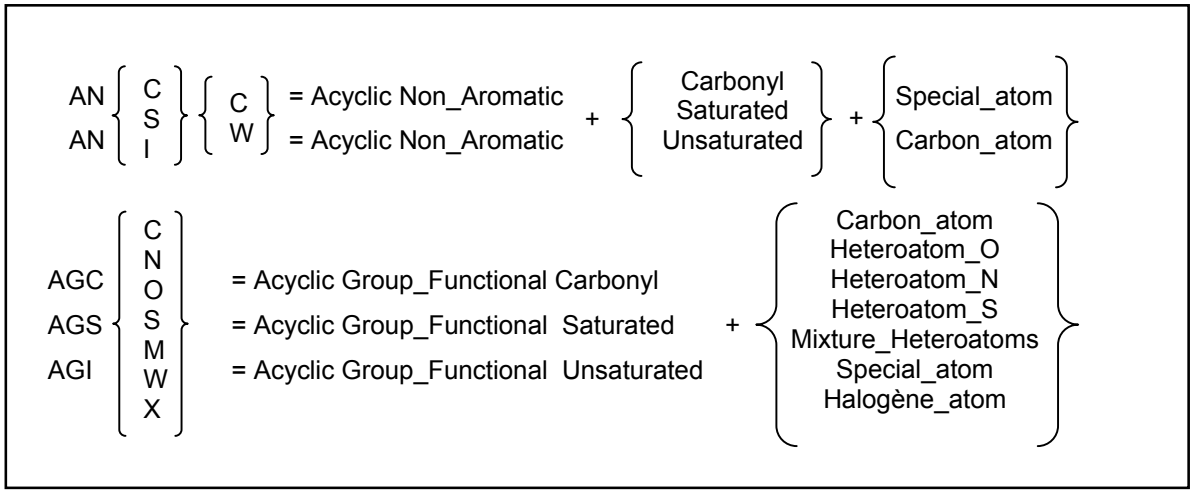


Figure II.5. Combinaisons possibles de noms de fichier de fragments acycliques dans FragDB

Ces données portent la taille de FragDB aux environs de 570 fragments. Les figures suivantes donnent des exemples de molécules appartenant aux bases décrites plus haut.

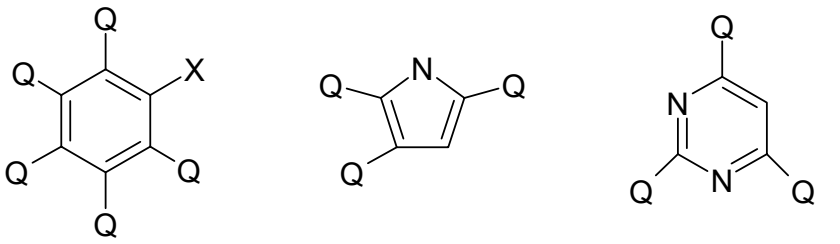


Figure II.6. Exemples de fragments CA (fragments contenant des cycles aromatiques)

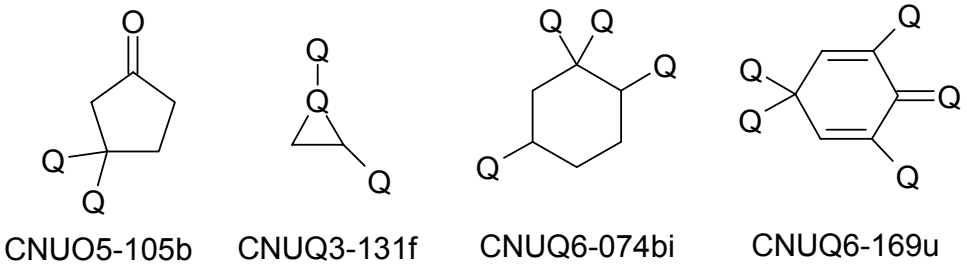


Figure II.7. Exemples de fragments CN (fragments contenant des cycles non aromatiques)

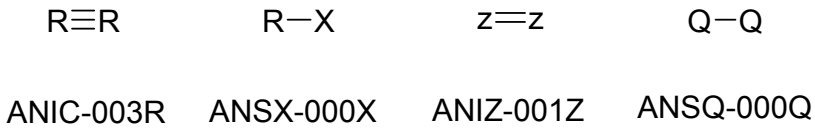


Figure II.8. Exemples de fragments AN (fragments contenant des chaînes acycliques)

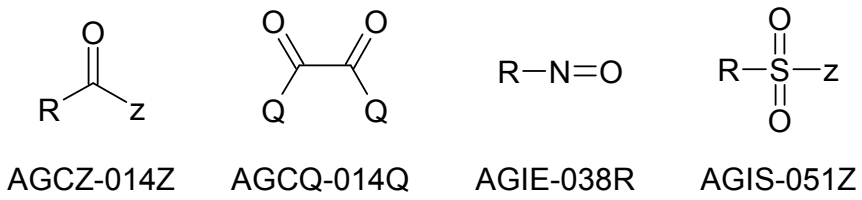


Figure II.9. Exemples de fragments AG (fragments contenant des groupes fonctionnels)

II.1.3 Construction des bases de données QueryDB et TestDB

Le groupe de molécules qui présentent des caractéristiques intéressantes pour l'usager et qui serviront des cibles pour les analyse de (dis)similarité, est appelé « QueryDB ». Le groupe de molécules à être testé, et sur lequel en cherche des resemblances avec la(les) cible(s) est appelé « TestDB ».

L'introduction des bases « QueryDB » et « TestDB » est faite par l'usager à l'aide d'une interface graphique. Il devra ensuite définir le nombre de cibles et des molécules test, ainsi que le type d'analyse à effectuer (similarité ou diversité). Pour l'acquisition des données chimiques, toutes les molécules devront être définies dans un format MOL valide, comme a été indiqué dans la section II.1.1.

D'autres limitations de « QueryDB » et de « TestDB » sont présentées dans la figure II.10. Ces valeurs (modifiables) répondent à un compromis entre la nécessaire optimisation de l'outil informatique et la volonté de couvrir un maximum des cas.

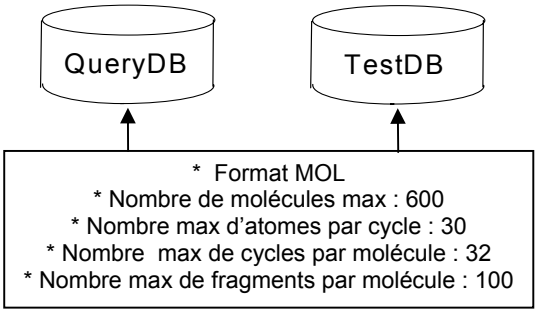

QueryDB
TestDB
* Format MOL
* Nombre de molécules max : 600
* Nombre max d'atomes par cycle : 30
* Nombre max de cycles par molécule : 32
* Nombre max de fragments par molécule : 100

Figure II.10. Restrictions des bases « QueryDB » et « TestDB »

Les modes de stockage en mémoire qui sont utilisés dans le traitement des informations structurales chimiques se différencient selon les applications, les algorithmes utilisés et l'architecture des ordinateurs.

Les modes que nous avons adoptés ont été choisis en fonction des applications, notamment le criblage de haut débit. Ces formats doivent être bien adaptés pour préserver à long terme les informations et pour échanger plus facilement les données.

II.2. Structuration des informations moléculaires et XML

Dans la section II.1.1 et II.1.2 ont été posées les bases de la construction de la base de sous-structures moléculaires nécessaires pour notre outil de criblage virtuel. Les informations manipulées jusqu'à maintenant concernent les fichiers MOL, les noms de fichiers, et des informations de nature physicochimiques pour compléter la description de la molécule à partir de ses fragments fondamentaux. Différents niveaux de complexité de l'information devront donc être intégrés pour optimiser la structuration et minimiser la redondance dans notre base de données.

La recherche d'une méthode simple, extensible et standard pour structurer l'information contenue dans notre base de données a abouti à l'utilisation des langages de marquage (XML).

II.2.1 Les langages de marquage

II.2.1.1 Historique [Murray-Rust2002]

Les origines de XML (langage de balisage extensible) remontent aux années 60 avec l'introduction par IBM de GML et son standard SGML. Ces deux langages permettaient de formater les documents texte et de définir leur type. Leur complexité d'implémentation a restreint leur utilisation à la communauté des éditeurs. Dans les années 90, l'apparition de HTML a permis la popularisation du web et de la

présentation informatisée de documents. Ce langage simple et facile à implémenter, a facilité l'échange et la présentation des contenus mais avec la contrainte d'être fixe, prédéfini et non modifiable.

En 1998 le W3C (*World Wide Web Consorsium*) recommande l'usage de XML qui devrait avoir comme objectifs :

- Pouvoir être utilisé sans difficulté sur Internet ;
- Soutenir une grande variété d'applications ;
- Etre compatible avec SGML ;
- Permettre de créer facilement des documents XML ;
- Permettre d'écrire facilement des programmes traitant les documents XML ;
- Permettre de produire des documents lisibles par l'homme et raisonnablement clairs ;
- Avoir une conception formelle et concise ;

Le XML est donc un meta-langage qui permet de représenter et de structurer l'information, en reprenant l'idée initiale de SGML mais en adoptant la simplicité de HTML.

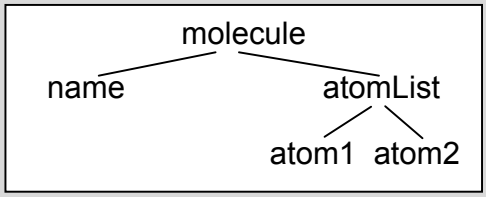
II.2.1.2 Principes [RecomXMLWeb]

Chaque document XML contient un ou plusieurs éléments, dont les limites sont marquées soit par des balises `<ouvrantes>` et `</fermantes>`, soit, par une balise d'élément `<vide/>`. L'information se trouve ainsi encapsulée dans des balises, ce qui rend plus facile la recherche et l'analyse d'éléments par un programme ou une personne. Les éléments de XML sont extensibles (on peut en définir tant qu'on veut) et ont des relations entre eux (sous la forme d'arbres parents-fils).

Dans le code II.1, l'élément père `<molecule>` contient deux éléments fils : `<name>` et `<atomsList>`. L'élément `<atomsList>` contient lui même deux éléments `<atom>`. L'information correspondant à la molécule HCl est maintenant organisée d'une manière logique. Ainsi, cette manière

d’exprimer l’information, la rend compréhensible pour une personne et facile à retrouver pour un programme ou une unité logique.

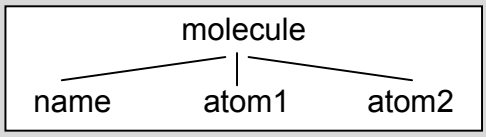
```
<molecule>
      <name> Hydrochloric acid </name>
      <atomList>
            <atom1> H </atom1>
            <atom2> Cl </atom2>
      </atomList>
</molecule>
```



Code II.1. Exemple de document XML

Nous avons pu choisir une autre manière d’ordonner l’information selon nos besoins. En tout cas l’information sera toujours structurée, non pas seulement présenté. Une possibilité alternative est montrée dans le Code II.2

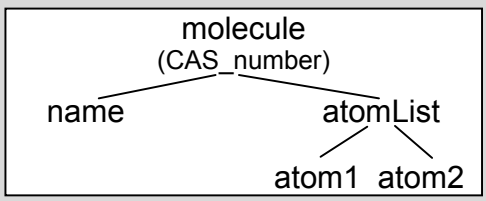
```
<molecule>
      <name> Hydrochloric acid </name>
      <atom1> H </atom1>
      <atom2> Cl </atom2>
</molecule>
```



Code II.2. Exemple de structure XML alternative

De plus, chaque élément peut avoir une paire attribut / valeur. Les attributs sont utilisés pour donner des informations additionnelles aux éléments qui structurent les données. Par exemple, dans le cas précédent, il se peut que le nombre CAS de HCl soit important pour une application particulière. Il sera alors représenté comme un attribut de l’élément `<molecule>`. Sa valeur `(7647-01-0)` devra être enfermée entre guillemets (“ ”) pour être reconnaissable.

```
<molecule CAS_number=“7647-01-0”/>
      <name> Hydrochloric acid </name>
      <atomList>
            <atom1> H </atom1>
            <atom2> Cl </atom2>
      </atomList>
</molecule>
```



Code II.3. Exemple de document XML avec la présence d’attributs et de valeurs

Pour interpréter correctement les données structurées avec XML, il est nécessaire de respecter la grammaire décrite précédemment. Ainsi les règles de liaison, d'ordre et de combinaison des balises sont spécifiées par la Définition de Type de Document (DTD). La DTD a pour but de définir chaque élément en précisant son contenu (comme une expression régulière introduisant la séquence (,) ou l'alternative (|) d'un nombre d'autres éléments) et ses attributs (en précisant le type de valeur prise, la présence exigée ou optionnelle et éventuellement la valeur par défaut).

```
<!-- Document Type Definition for the code I.3 --!>

<!ELEMENT molecule (name,atomList)>
<!ATTLIST molecule CAS_number CDATA #REQUIRED>

<!ELEMENT atomList (atom1,atom2)>
<!ELEMENT atom1 (#PCDATA)>
<!ELEMENT atom2 (#PCDATA)>
```

Code II.4. Exemple de DTD

Cette DTD déclare les trois types d'éléments présents dans le code II.3 : `<molecule>`, `<name>` et `<atomsList>`. Chacun de ces éléments est défini par son contenu à l'aide du mot clé !ELEMENT, et par ses attributs avec le mot clé !ATTLIST. Dans notre exemple de DTD on observe deux types de données : attributs de type chaîne de caractères (CDATA) obligatoire (#REQUIRED) et éléments de type chaîne de caractères (#PCDATA).

Le document XML présenté dans le code I.3 est un document « valide » au sens XML vis-à-vis de la DTD du code I.4, car la grammaire y est conforme à celle définie par la DTD. Cette notion de validité était déjà présente dans SGML, mais la norme XML ajoute une nouvelle notion (moins forte) qui est celle de document « bien formé ». Un document est dit bien formé si les balises qui le composent forment un et un seul arbre, ce qui est bien entendu une condition nécessaire à sa validité. Ainsi, les

documents XML peuvent être manipulés indépendamment de leur DTD. Cette particularité est même une des motivations qui a présidé à son élaboration, le faisant passer d'un langage documentaire à un langage d'échange de données structurées. Une conséquence directe de cette notion est qu'elle a favorisé l'apparition d'autres langages pour exprimer la structure des documents et des données. Nous évoquerons plus tard le Langage de Marquage Chimique (CML).

Il existe d'autres types de contrôle et de validation de documents XML : Le *XMLSchema* (une puissante extension des DTD en XML) les *namespaces* (évite la collision des noms et élimine l'ambiguïté), etc. L'interrogation de bases de données (*XQuery*) et la transformation de documents XML (XSLT) font partie des taches qui ont été développées par le W3C depuis l'apparition des langages de marquage. Dans la figure suivante nous avons regroupé quelques éléments de construction de documents XML ainsi que des applications courantes, notamment en sciences, en documentation multimédia et dans le Web Sémantique.

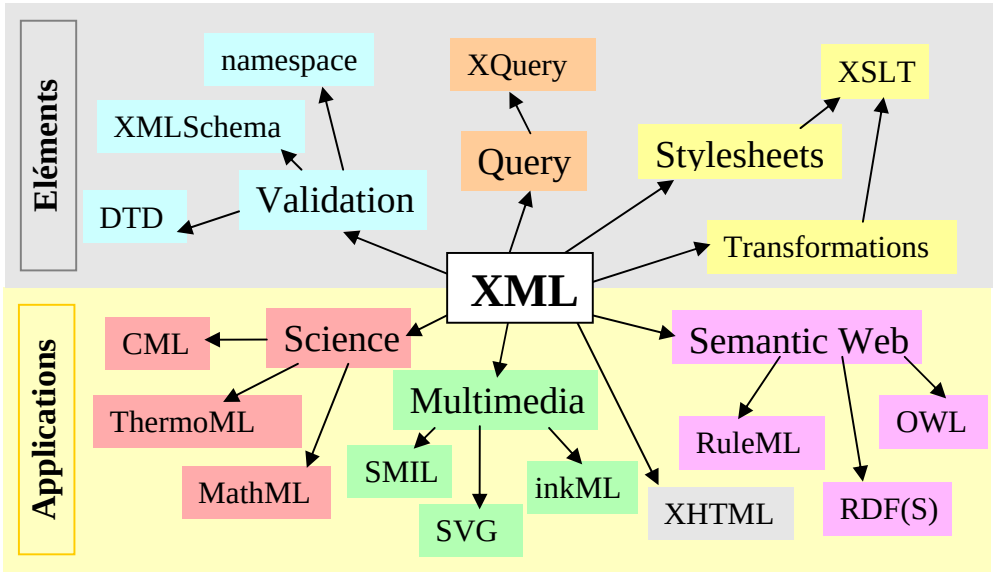


Figure II.11. La famille XML (adapté de [Bolev2001])

Quelques sigles de la figure II.11 ont déjà été expliqués lors de l'introduction au langage XML tout au long de la section présente, toutefois on peut trouver le reste dans la section des abréviations au début du manuscrit.

De nombreux livres [Harold2001] et tutoriaux [TutorialXMLWeb] sont disponibles pour apprendre à utiliser XML. Toutefois la recommandation du W3C [RecomXMLWeb] reste le document de référence.

II.2.1.3 XML pour structurer les informations chimiques

Depuis l'apparition des langages de marquage, beaucoup d'efforts ont été faits dans différents champs scientifiques pour définir des schémas et des vocabulaires ainsi que des ontologies, regroupant les connaissances actuelles du domaine. Il est important de noter que pour le domaine particulier de la chimie, la construction d'un langage de marquage a été l'une des priorités des groupes de travail du W3C [Murray-Rust2002, Gkoutos2001].

Les résultats des ces efforts ont abouti à la création d'une base extensible pour un langage de marquage chimiquement compatible appelé CML [CMLWeb]. CML représente une des premières approches pour traiter la plupart des problèmes d'échange d'information chimique à travers le Web et autres réseaux [Murray-Rust1999, Murray-Rust2001, Murray-Rust2001a, Murray-Rust2003]. Ce langage permet à l'usager de structurer dans un cadre commun, l'information chimique déjà extraite, analysée, partagée ou visualisée.

L'implémentation de XML dans un cadre chimique a été en partie facilitée par la création de CML et l'utilisation par des entreprises et des universités des langages de marquage comme format d'échange d'information.

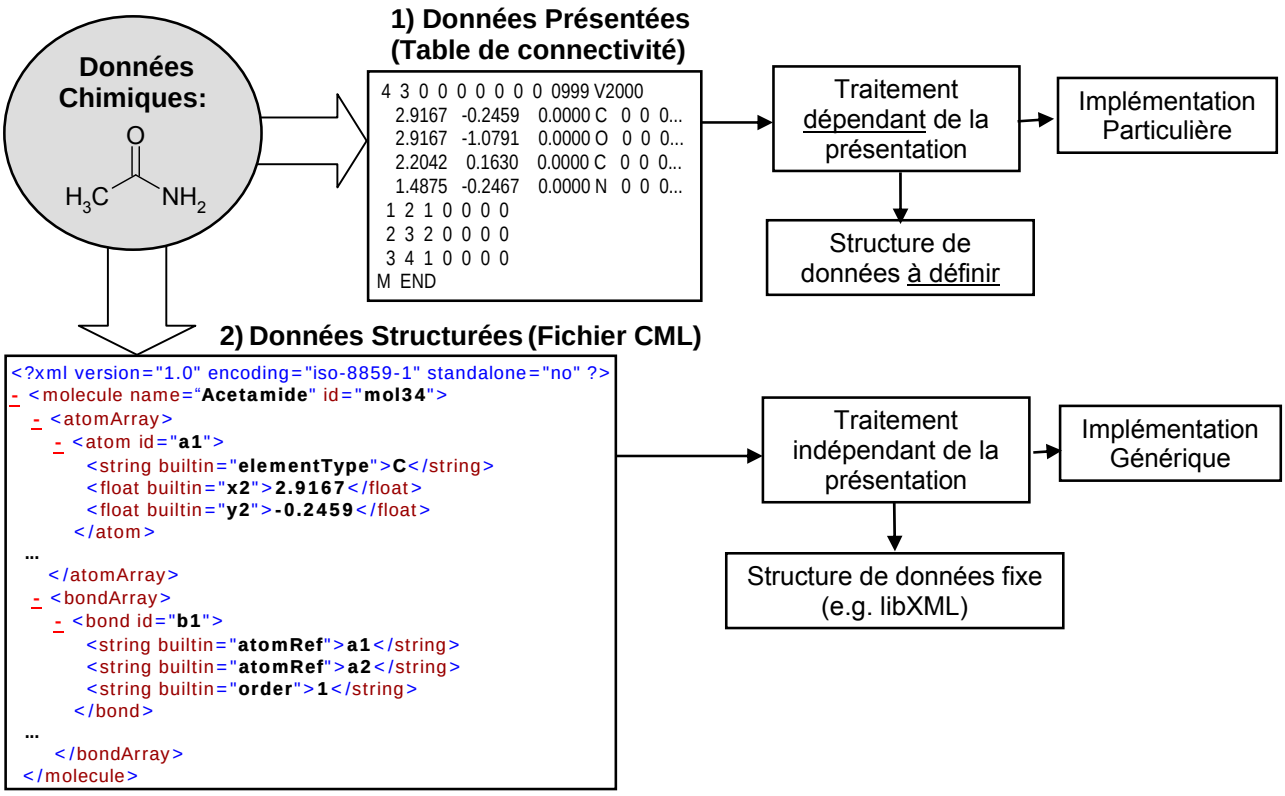


Figure II.12. Deux cas d'étude pour la structuration de l'information chimique.

Dans notre cas particulier, deux cas d'étude ont été proposés pour la structuration des informations chimiques contenues dans un fichier .MOL des molécules (voir figure II.12).

1) Table de connectivité :

- Information dépendante de la présentation (un changement de la table de connectivité rendra le fichier invalide pour la lecture).

- Structure de données dépendant de la présentation et à définir par l'usager.

- Implémentations limitées au cadre des données présentées.

2) Fichier XML :

- Information indépendante de la présentation (un rajout ou une modification du fichier XML n'intervient pas dans la lecture).
- Structure de données fixe et indépendant de la présentation.
- Possibilité d'utiliser plusieurs structures de données sans changer l'implémentation.

Ainsi, au moment de créer notre base de données de sous-structures pour notre outil de criblage virtuel, il n'a pas été nécessaire de prévoir à l'avance toutes les possibilités des futures implémentations de l'information, grâce au cadre flexible et extensible de XML. Ceci permet d'effectuer facilement des modifications et des additions d'information sans obligation de modifier la structure des données.
On obtient ainsi une génération automatique (et dynamique) de structures de données par extraction de l'information structural. Le programme traitera tout ce que l'utilisateur lui donnera en forme de données chimiques (aromaticité, cycles hétéoatomiques, etc) sous condition que la molécule soit dans un format MOL valide.

Une fois les traitements sur les molécules effectués, on peut mettre en forme les résultats de manière automatique avec une feuille de style XSLT, qui interprète les informations enfermées dans les balises XML et qui les affiche sous forme de tableaux, de texte, etc. La figure II.12 résume les avantages du format XML par rapport aux formats conventionnels.

Dans la section suivante, nous aborderons la représentation et la structuration des données chimiques de nature mixte qui compose notre base de sous-structures chimiques. Des exemples de l'implémentation du code XML pour construire un index de fragments ainsi que les DTDs correspondants, seront également donnés.

II.2.2 La structuration de la FragDB avec XML

Dans cette section nous allons expliquer en détail la manière dont nous avons représenté et structuré la base de fragments FragDB. De la création de l'index des fragments jusqu'à son remplissage et sa validation, l'utilisation de XML comme format de données pour structurer les informations moléculaires complexes montre beaucoup d'avantages.

II.2.2.1 Création et remplissage d'un index-XML de fragments

Dans la section II.1 nous avons détaillé les principes de construction de la base de données FragDB : le choix des fragments, des atomes génériques, la « nomenclature » du nom de fichier, etc. Dans cette section nous sommes concernés par la base de fragments, FragDB illustrée dans la figure suivante montrant le lexique utilisé pour désigner les bases des molécules utilisées dans l'outil de criblage.

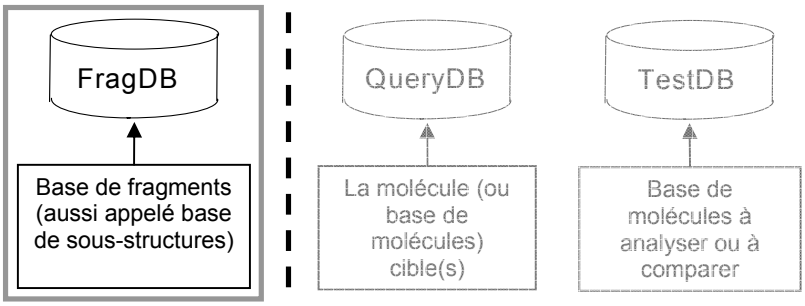

FragDB
Base de fragments (aussi appelé base de sous-structures)
QueryDB
La molécule (ou base de molécules) cible(s)
TestDB
Base de molécules à analyser ou à comparer

Figure II.13. Nous centrons notre attention sur la base de fragments/sous-structures (FragDB)

La FragDB est à l'origine un groupe de fichiers MOL nommés selon une « nomenclature » particulière et construits suivant certaines règles, d'atomes génériques et de fréquence des sous-structures. La base de données ainsi construite est formée de fragments isolés et non ordonnés, sans aucune priorité, et enfin, non structurés. L'extraction des informations chimiques, dans ces conditions est particulièrement difficile et hasardeuse.

Nous avons donc décidé de créer une structure de données qui nous permet d'exploiter les informations chimiques contenues dans le fichier MOL. Nous avons appelé cette structure de données : index-XML.

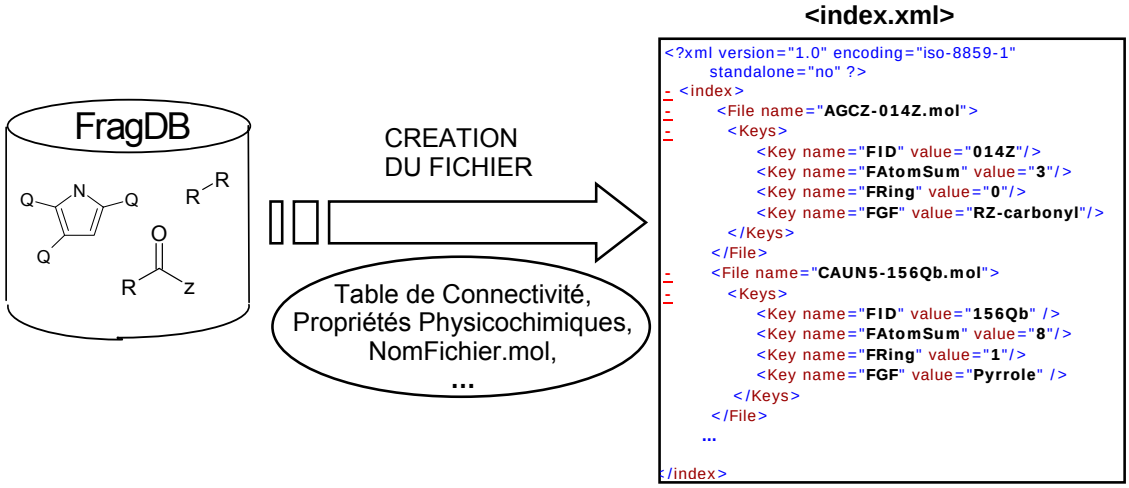


Figure II.14. Création et remplissage d'un index.xml à partir de la FragDB.

L'index-XML est composée d'autant éléments `<file>` qu'il a de fragments. Chaque élément `<file>` contient plusieurs sous-éléments : `<Keys>, <Properties>`. La figure II.14 montre les étapes de la création et du remplissage d'un index-XML.

On note sur la figure que les deux derniers sous-éléments ne sont pas montrés (pour des raisons de clarté), toutefois il est indiqué le contenu de la balise `<Keys>` pour deux des fragments montrés dans la FragDB : AGCZ-014Z.mol et CAUN5-156Qb.mol.

Il est important de noter la structure indexée des données XML en forme d'arbre, ce qui facilite la lecture et la compréhension des contenus, autant pour l'homme que pour les outils d'extraction ou d'interrogation d'information. Les noms et les nombres des éléments peuvent être modifiés à tout moment sans altérer le traitement de la FragDB par des outils nommés préalablement.

L’index-XML contient entre autres les informations suivantes :

- Des pointeurs vers les fichiers MOL de la FragDB,
- Les informations moléculaires extraites du nom de fichiers,
- Des clefs de recherche,
- Des propriétés physicochimiques des fragments,
- Et toute autre information susceptible de compléter la FragDB.

L’index de sous-structures est généré automatiquement à partir de la FragDB en suivant un algorithme simple en langage C. La figure II.15 montre un aperçu de l’algorithme. Le fichier est ensuite rempli en mémoire (figure II.16) en utilisant les informations disponibles (figure II.17).

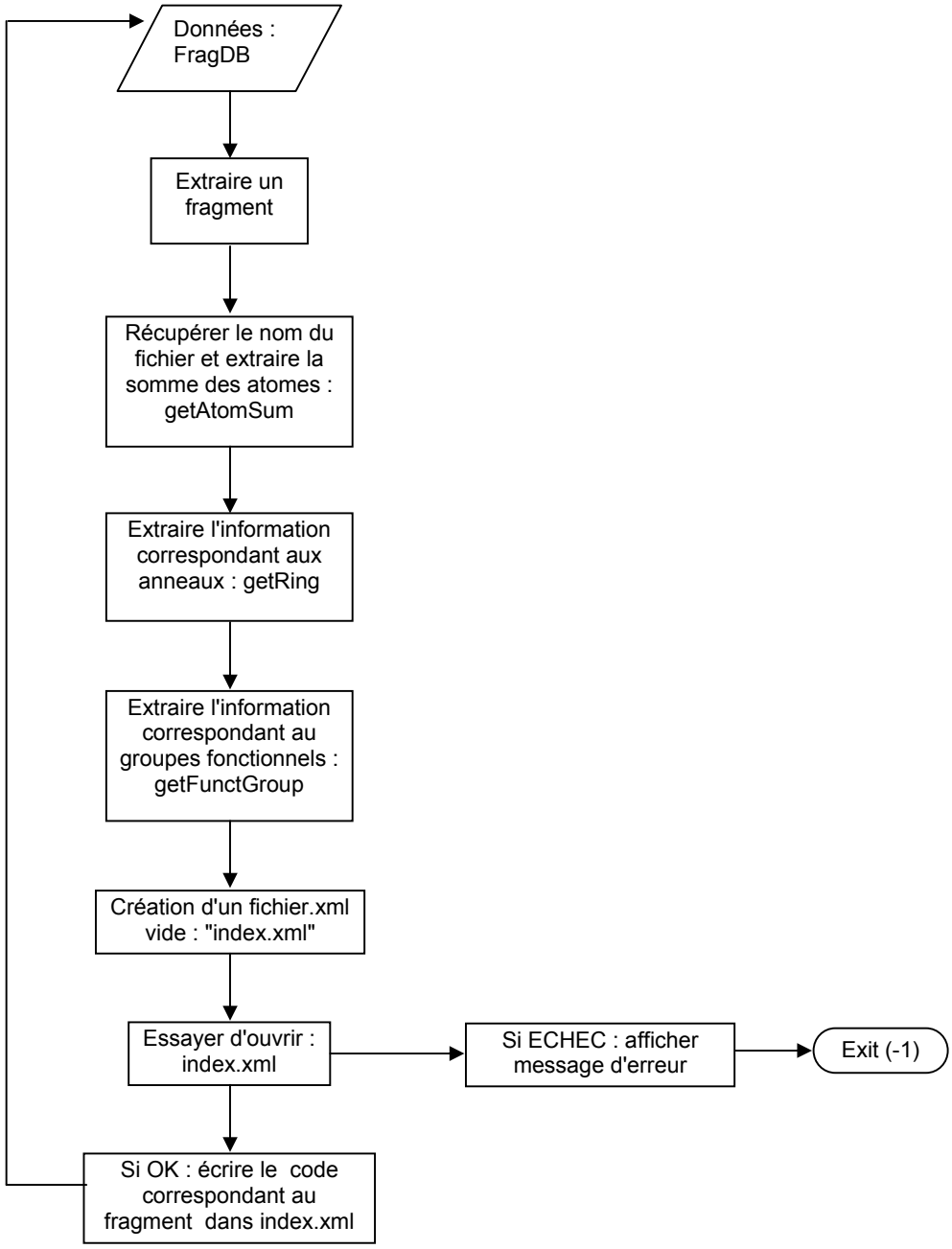

Données :
FragDB
Extraire un fragment
Récupérer le nom du fichier et extraire la somme des atomes : getAtomSum
Extraire l'information correspondant aux anneaux : getRing
Extraire l'information correspondant au groupes fonctionnels : getFunctGroup
Création d'un fichier.xml vide : "index.xml"
Essayer d'ouvrir : index.xml
Si ECHEC : afficher message d'erreur
Exit (-1)
Si OK : écrire le code correspondant au fragment dans index.xml

Figure II.15. Algorithme pour la création d'un index-XML de fragments, à partir d'une base de données.

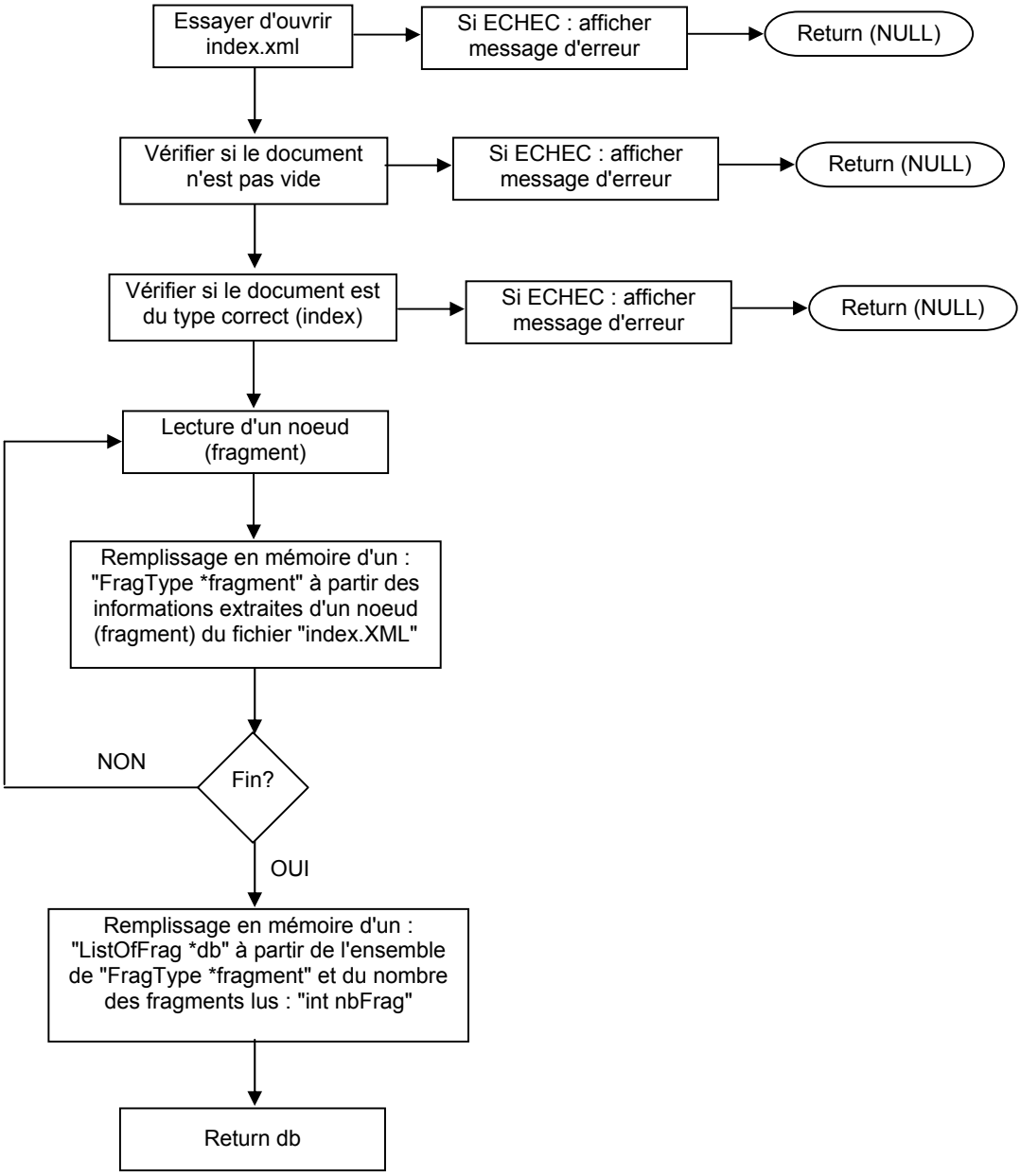

Essayer d'ouvrir index.xml
Si ECHEC : afficher message d'erreur
Return (NULL)
Vérifier si le document n'est pas vide
Si ECHEC : afficher message d'erreur
Return (NULL)
Vérifier si le document est du type correct (index)
Si ECHEC : afficher message d'erreur
Return (NULL)
Lecture d'un noeud (fragment)
Remplissage en mémoire d'un : "FragType *fragment" à partir des informations extraites d'un noeud (fragment) du fichier "index.XML"
NON
Fin?
OUI
Remplissage en mémoire d'un : "ListOfFrag *db" à partir de l'ensemble de "FragType *fragment" et du nombre des fragments lus : "int nbFrag"
Return db

Figure II.16. Algorithme pour l'ouverture et remplissage en mémoire d'index-XML

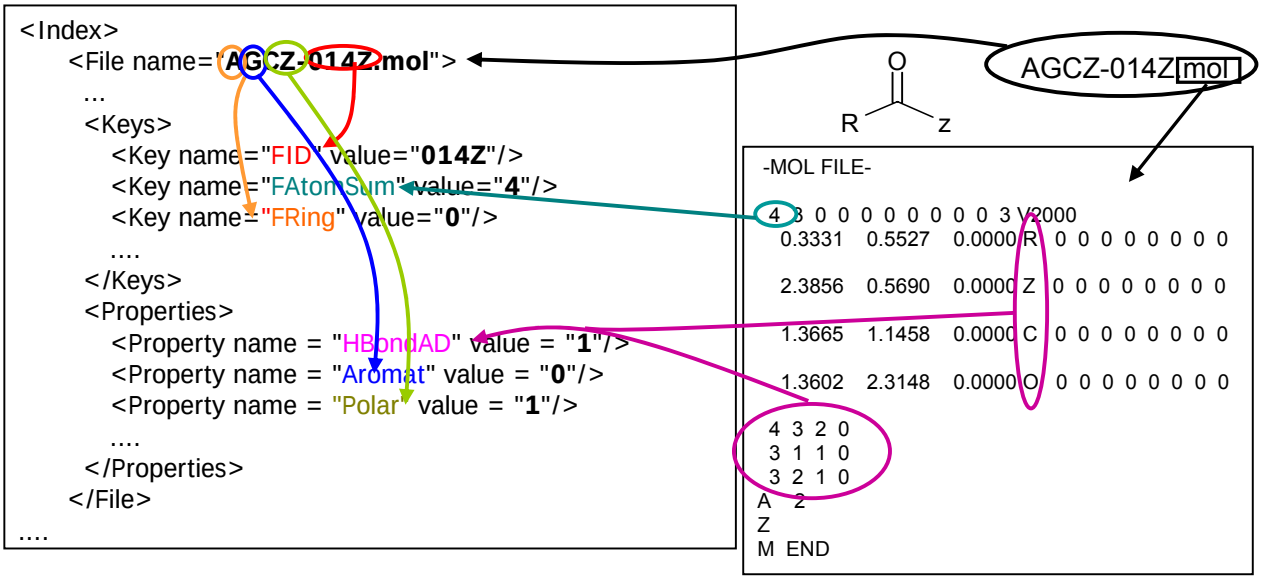


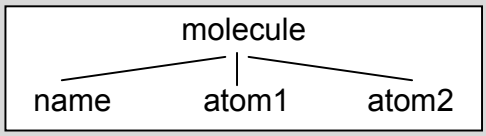
Figure II.17. Remplissage de la structure de données en utilisant des informations extraites à partir du nom de fichier (voir figure I.3) et du fichier .MOL (voir annexe 3).

La création d'un fichier XML pour structurer des données chimiques complexes est un processus simple et rapide et qui peut être effectué automatiquement. Le langage XML comme tous les métalangages permet de définir ses propres éléments et donc de s'adapter à chaque domaine (chimie médicinale, chimie inorganique, spectroscopie, etc.). Le langage est flexible et extensible, et les informations plus faciles à retrouver automatiquement car elles sont « enfermées » dans les éléments.

Toutefois deux inconvénients sont à noter : Les fichiers XML générés sont d'une taille assez grande, car le langage a besoin de beaucoup de texte pour décrire des informations parfois simples. Dans l'annexe 2, à la fin du manuscrit est inclus un fichier index-XML qui occupe plusieurs pages malgré le fait qu'il ne contienne qu'une quantité restreinte des fragments. Récemment, des fichiers binaires pour XML ont été proposés comme alternative aux fichiers conventionnels, ce qui réduit considérablement l'encombrement [BinXML]. Le deuxième inconvénient est lié à la nature même des langages de marquage : malgré le fait d'avoir des informations très bien structurées, un robot ou logiciel ne

comprendra pas leur sens. Par exemple, dans le code suivant, l'élément <molécule> se réfère clairement à la molécule de HCl composée d'un atome d'hydrogène et d'un atome de chlore, ceci est assez compréhensible pour un humain.

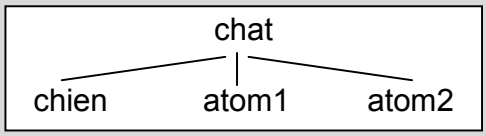
```
<molecule>
      <name> Hydrochloric acid </name>
      <atom1> H </atom1>
      <atom2> Cl </atom2>
</molecule>
```



Code II.5. Exemple de document XML

Si nous échangeons l'élément `<molecule>` et `<name>` par `<chat>` et `<chien>`, nous obtenons le code II.6. Ce document XML est parfaitement valable car il respecte les règles de syntaxe et de grammaire d'XML fixés pour la DTD, mais en même temps il n'a aucun sens chimique. On pourrait même interroger le document on lui demandant l'élément `<chien>` et au retour on aura la chaîne de caractères « Hydrochloric acid ».

```
<chat>
      <chien> Hydrochloric acid </name>
      <atom1> H </atom1>
      <atom2> Cl </atom2>
</chat>
```



Code II.6. Document XML modifié

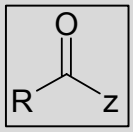
Cet inconvénient nous amène à la prochaine étape dans la structuration de données avec les langages de marquage : la représentation des connaissances. Ceci sera le sujet de la section II.2.5

Dans la prochaine partie nous aborderons l'outil qui permet de valider notre document XML pour son futur traitement ou échange : la Définition de Type de Document ou DTD.

II.2.2.2 Une DTD pour valider l'index-XML

La DTD a pour but de définir les règles de liaison, d'ordre et de combinaison des balises dans un document XML [DTDWeb]. Ceci permet notamment de bien interpréter les données structurées avec XML et d'éviter des erreurs de syntaxe ou de grammaire qui auraient pu s'infiltrer dans l'édition du document. Nous avons déjà expliqué la manière de construire une DTD et la signification des termes la composant (section 2.1.2).

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<!-- Sample of Index.xml : Data Structure for FragDB  -->
<Index>
      <File name="AGCZ-014Z.mol">
            <Keys>
                  <Key name="FID" value="014Z"/>
                  <Key name="FAtomSum" value="4"/>
                  ....
            </Keys>
            <Properties>
                  <Property name = "HBondAD" value = "1"/>
                  <Property name = "Aromat" value = "0"/>
                  <Property name = "Polar" value = "1"/>
                  ....
            </Properties>
      </File>
....
</Index>
```



Code II.7. Index-XML : Structure de données simplifiée pour FragDB

L'implémentation de l'index-XML a nécessité la définition préalable d'une DTD correspondante.

Dans le code II.7 on présente un fragment du fichier index-XML pour la sous-structure AGCZ-014Z.mol déjà apparue dans la figure II.17. L'information contenue dans l'index est reprise ici dans une version simplifiée pour des questions de place.

Par la suite, dans le code II.8 et en suivant les règles de construction, nous avons proposé une DTD pour la validation de la version simplifiée du fichier index-XML. On peut noter la définition des éléments et de ses attributs, ainsi que du type des données composant l'index-XML. Une version non

simplifiée des structures de données, des DTDs, ainsi que de l'index de fragments a été inclus dans l'annexe 2.

```
<!-- Sample of DTD for index.xml -->
<!ELEMENT index (File+)>
<!ELEMENT File (Keys,Properties)>
<!ATTLIST File name CDATA #REQUIRED>
<!ELEMENT Keys (Key+)>
<!ATTLIST Key name CDATA #REQUIRED>
<!ATTLIST Key value CDATA #REQUIRED>
...
<!ELEMENT Properties (Property+)>
<!ATTLIST Property name CDATA #REQUIRED>
<!ATTLIST Property value CDATA #REQUIRED>
...
```

Code II.8. DTD simplifiée pour l'Index-XML

II.2.3 La structuration du QueryDB et du TestDB

Dans cette section nous allons expliquer en détail la manière dont nous avons obtenu, représenté et structuré les molécules appartenant au QueryDB et au TestDB. Nous montrerons également les structures de données nécessaires pour l'extraction et le traitement des composants.

II.2.3.1 Transformation des molécules et création du VecteurRepresentatif-XML

Dans la section II.2.2 nous avons détaillé les principes de construction et de structuration de la base de données FragDB. Maintenant nous sommes concernés par la base de molécules cibles (QueryDB) et les molécules à comparer ou test (TestDB). Ceci est illustré dans la figure II.18.

Ces molécules seront définies et introduites dans l'outil par l'usager. Toutefois il faut veiller à ce que certaines conditions soient remplies :

- Tous les fichiers doivent être en format MOL ;

- Les atomes C des sous-structures aromatiques doivent avoir comme type de liaison 4 (option par défaut quand on construit les molécules avec des liaisons aromatiques en pointillés et non par alternance de doubles et simples liaisons) ;
- Les molécules doivent être bien définies (donc respectant les lois chimiques) ;
- En règle générale, l'outil retiendra ce que l'usager a écrit sur la molécule.

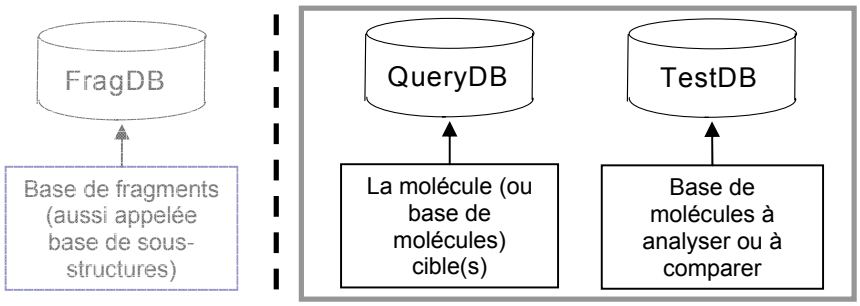

FragDB
Base de fragments (aussi appelée base de sous-structures)
QueryDB
TestDB
La molécule (ou base de molécules) cible(s)
Base de molécules à analyser ou à comparer

Figure II.18. Nous centrons notre attention sur la base de données moléculaires

Toutes les molécules du QueryDB et du TestDB subissent une transformation pour extraire leurs informations et construire des vecteurs contenant les données nécessaires pour la comparaison des molécules.

Le détail du processus de transformation fera partie des sujets traités au chapitre III. Pour l'instant nous nous intéressons à leur structure et à son organisation dans un fichier XML.

Dans la figure II.19 la création d'un index des molécules est représentée. Pour calculer les similarités entre molécules on doit préalablement avoir transformé les molécules à analyser. Les descripteurs sont générés par comparaison atome-atome entre les molécules de la QueryDB-TestDB et les fragments de la FragDB (voir chapitre III). Une fois que chaque molécule a sa représentation bien définie, on passe à la structuration de cette information.

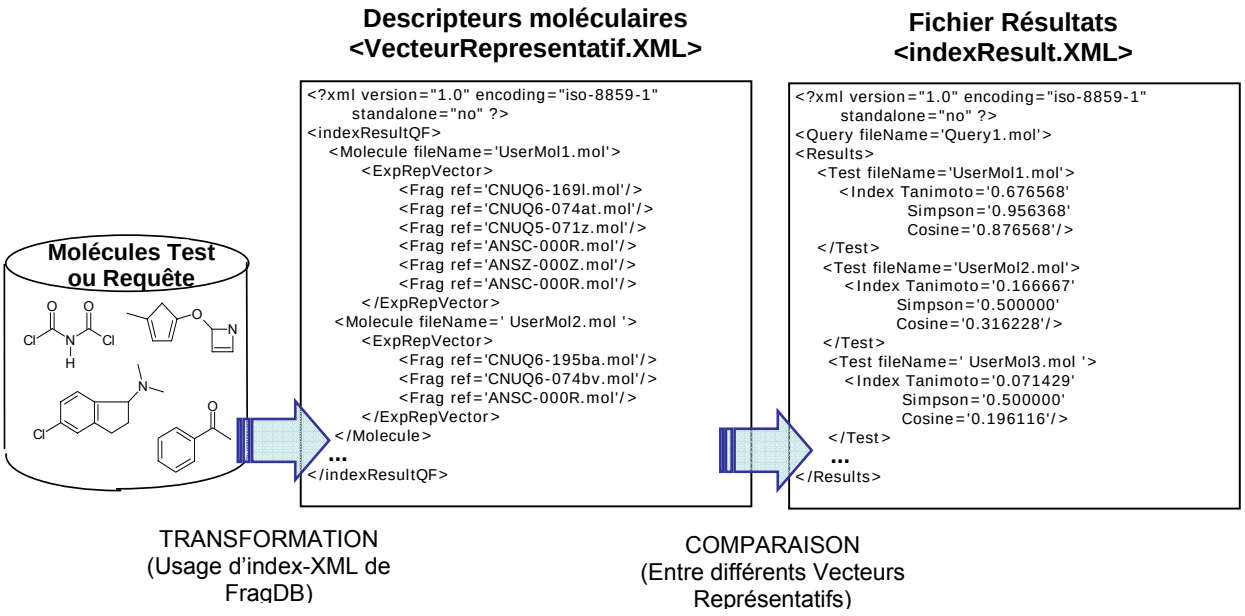


Figure II.19. Création et remplissage d'un indexResult-XML à partir des molécules de la QueryDB-TestDB. Une étape intermédiaire importante est la transformation des molécules à analyser dans une représentation vectorielle des fragments.

Pour cela on utilise à nouveau les avantages des fichiers XML par rapport aux bases de données conventionnelles. En plus, la vocation HTS (*High Thoughput Screening*) de notre outil nous oblige à être capables de travailler avec des bases de données moléculaires de grande taille. Comme conséquence, la structure proposée doit être suffisamment flexible pour accepter des modifications ou des extensions, sans modification drastique des traitements effectués sur les informations moléculaires. Le code II.9 reprend l'exemple montré dans la figure II.19. On observe que le VecteurRepresentatif-XML est composée d'autant éléments `<Molecule>` qu'il y a de molécules dans la base moléculaire TestDB. Chaque élément `<Molecule>` a comme attribut le nom de la molécule analysée, et contient un sous-élément : `<ExpRepVector>`. Ces éléments contiennent eux-mêmes une liste de `<Frag>` qui ont comme attribut le nom du fragment correspondant. Pour finir l'élément `<indexResultQF>` regroupe la liste des listes nommées ci-dessus.

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<indexResultQF>
    <Molecule fileName='UserMol1.mol'>
          <ExpRepVector>
              <Frag ref='CNUQ6-169l.mol'/>
              <Frag ref='CNUQ6-074at.mol'/>
              <Frag ref='CNUQ5-071z.mol'/>
              <Frag ref='ANSC-000R.mol'/>
              <Frag ref='ANSZ-000Z.mol'/>
              <Frag ref='ANSC-000R.mol'/>
          </ExpRepVector>
     </Molecule>
     <Molecule fileName=' UserMol2.mol '>
          <ExpRepVector>
              <Frag ref='CNUQ6-195ba.mol'/>
              <Frag ref='CNUQ6-074bv.mol'/>
              <Frag ref='ANSC-000R.mol'/>
          </ExpRepVector>
     </Molecule>
  ...
</indexResultQF>
```

Code II.9.Exemple de « VecteurRepresentatif-XML ». Deux molécules d'une TestDB donnée sont analysées. Les fragments correspondants et ses informations attachées sont regroupés dans une liste, pour être comparés par la suite entre eux.

Le fichier XML ainsi généré et que l'on appelle « VecteurRepresentatif-XML » contient donc, des informations de nature mixte :

- Des pointeurs vers les sous-structures MOL de la FragDB,
- Des pointeurs vers les fichiers MOL de la QueryDB-TestDB,
- Des informations moléculaires extraites des noms de fichiers,
- Le nombre de molécules cible et test,
- Les informations des vecteurs attachés à chaque molécule (cible ou test),
- Des clefs de recherche,
- Et toute autre information susceptible d'aider à comparer QueryDB-TestDB.

Ce fichier est généré automatiquement et de manière récursive à partir des fragments (FragDB) et de molécules (QueryDB-TestDB) en suivant un algorithme simple (« Comparaison fragment-molécule » présenté dans la figure II.20).

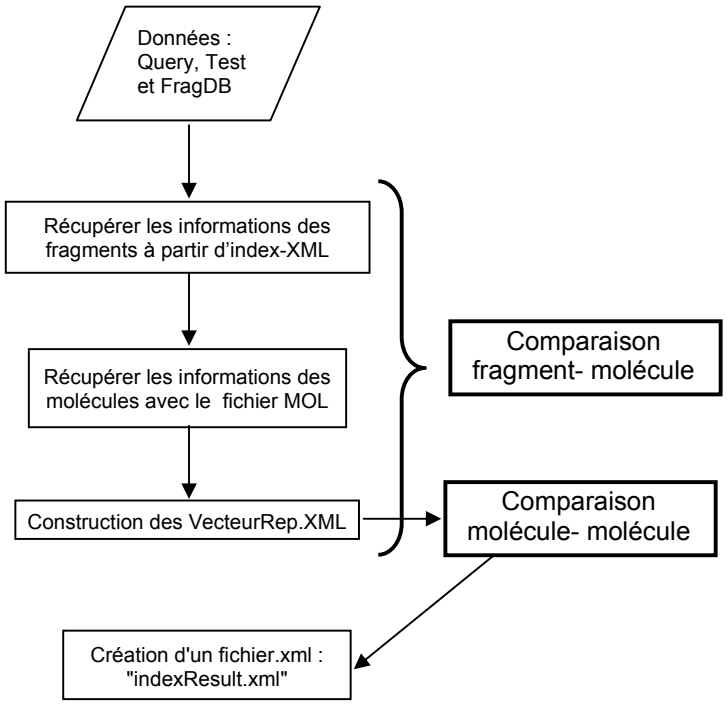


Figure II.20. Algorithme pour la création d'un index-XML de fragments, à partir d'une base de données.

Par comparaison des descripteurs moléculaires et l'usage des coefficients ou des distances de similarité / diversité, on peut effectuer l'analyse de la base. Ceci est représenté dans la figure II.20 « Comparaison molécule-molécule » et sera traité en détail dans le chapitre IV.

La création d'un VecteurRepresentatif-XML pour structurer des données moléculaires est un processus simple, rapide et effectué automatiquement. Dans la prochaine partie nous aborderons l'outil qui

permet de valider notre document XML pour le traitement ou l'échange futur du VecteurRepresentatif-XML : la Définition de Type de Document ou DTD.

II.2.3.2 Une DTD pour valider le VecteurRepresentatif-XML

Comme on a indiqué dans la section II.2.3.2, le *Document Type Definition* ou DTD a pour but de définir les règles de liaison, d'ordre et de combinaison des balises dans un document XML pour bien interpréter les données structurées avec XML.

Dans la section II.2.1.2 nous avons expliqué la manière de construire une DTD et la signification des termes la composant. Une DTD dépend étroitement du fichier XML auquel elle est rattachée.
Ainsi, l'implémentation du VecteurRepresentatif-XML montré dans le code II.9 a nécessité la définition préalable de sa DTD correspondante que nous montrons ci-dessus dans le code II.10. Cet exemple montre la structuration de deux molécules différentes et de leurs vecteurs correspondants. La DTD contient donc la définition des éléments et des attributs du modèle de VecteurRepresentatif-XML.

```
<!-- Sample of DTD -->
<!ELEMENT indexResultQF (Molecule+)>
<!ELEMENT Molecule (ExpRepVector)>
<!ATTLIST Molecule fileName CDATA #REQUIRED>
<!ELEMENT ExpRepVector (Frag+)>
<!ATTLIST Frag ref CDATA #REQUIRED >
```

Code II.10. DTD correspondant au fichier XML du code II.9

II.2.3.3 Une DTD pour valider l'indexResult-XML

Dans les figures II.19 et II.20, nous avons représenté d'une manière simplifiée, les deux étapes à suivre pour créer et remplir une liste d'indices de similarité et de diversité à partir de QueryDB-TestDB : la première étape étant la création des descripteurs moléculaires et la deuxième, la comparaison des

descripteurs pour obtenir des valeurs de similarité et de diversité indexés dans un fichier XML. Pour garantir la validité des résultats vis-à-vis le langage XML, il faut implémenter son correspondant DTD.

Le code II.11 reprend l'exemple montré dans la figure II.19. On observe que l'indexResult-XML est composée d'autant éléments `<Query>` qu'il y a de requêtes dans la QueryDB. Chaque élément `<Query>` a comme attribut le nom de la molécule requête, et contient un sous-élément appelé `<Results>`. Cet élément regroupe la liste de molécules `<Test>` qui ont été comparé avec la requête, ainsi que ses mesures de similarité sous la balise `<Index>`. La DTD correspondante est dans le code II.12.

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<Query fileName='Query1.mol'>
  <Results>
     <Test fileName='UserMol1.mol'>
        <Index Tanimoto='0.676568'
          Simpson='0.956368'
          Cosine='0.876568'/>
     </Test>
     <Test fileName='UserMol2.mol'>
         <Index Tanimoto='0.166667'
        Simpson='0.500000'
        Cosine='0.316228'/>
     </Test>
     <Test fileName='UserMol3.mol'>
         <Index Tanimoto='0.071429'
         Simpson='0.500000'
         Cosine='0.196116'/>
     </Test>
     ...
  </Results>
</Query>
```

Code II.11. IndexResult-XML pour une TestDB déterminé

```
<!-- Sample of DTD -->
<!ELEMENT Query (Results)>
<!ATTLIST Query fileName CDATA #REQUIRED>
<!ELEMENT Results (Test+)>
<!ATTLIST Test fileName CDATA #REQUIRED>
<!ELEMENT Test (Index)>
```

Code II.12. DTD correspondant au IndexResult-XML

II.2.4 La représentation des connaissances

Jusqu'à maintenant, nous avons montré comment les langages de marquage optimisaient la structuration des données en permettant une automatisation rapide et facile des processus d'interrogation et d'analyse des bases de données. Ainsi l'information chimique est « enfermée » et les mots « atome » et « molécule » deviennent manipulables par les machines.

Le problème est que parfois ceci n'est pas suffisant car les machines n'ont pas accès au sens de l'information manipulée. La représentation des connaissances intervient alors comme un moyen d'exprimer l'information et de la rendre compréhensible aux outils de traitement de données. Usuellement, le formalisme repose sur des langages logiques qui permettent la modélisation des ontologies, conceptualisant ainsi la connaissance du domaine (figure II.21).

Le terme ontologie, issu de la philosophie, désigne généralement l'ensemble des concepts d'un domaine. Dans le cadre de la représentation des connaissances, ce terme est employé plus particulièrement pour décrire les contenus du support: concepts, relations et contraintes qui sont effectivement utilisés pour modéliser un domaine donné. On peut considérer qu'une ontologie, dans ce sens, est l'aboutissement formel de la définition d'une terminologie.

Dans le contexte chimique, les ontologies regroupent un ensemble de définitions lisibles par des machines, qui créent une taxonomie de classes, des relations et des axiomes logiques [OWLWeb] définissant les règles des atomes, molécules, réactions, etc. En chimie, il y a un besoin croissant des ontologies. Celles-ci doivent couvrir l'information chimique indispensable pour la formalisation des concepts, ainsi que faciliter l'échange et la compression des processus.

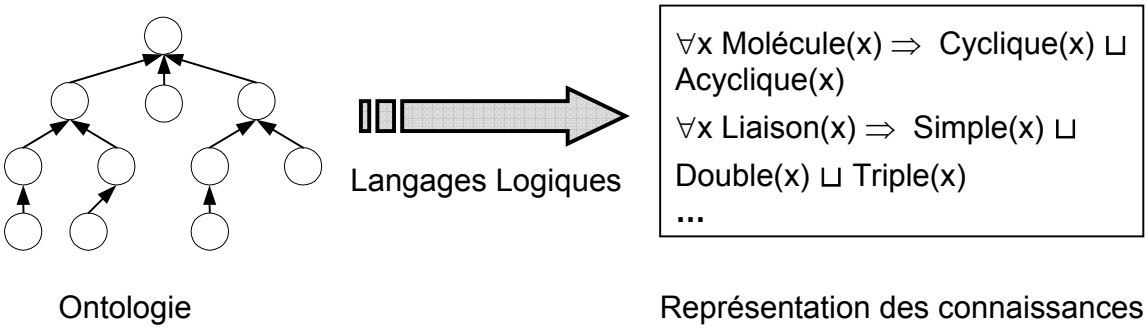


Figure II.21. La représentation des connaissances comme produit des ontologies. L'ontologie regroupe une taxonomie de classes, des relations et des axiomes logiques qui sont ensuite « traduits » en utilisant des langages logiques pour devenir compréhensible pour les machines. Ceci est le principe même de « représentation des connaissances ». Dans l'exemple on énonce deux règles chimiques qui se traduisent ensuite en langage logique: « Pour toute molécule, soit elle est cyclique, soit acyclique » et « Pour toute liaison, soit elle est simple, soit double, soit triple, etc ».

Actuellement, aucune ontologie chimique n'est encore disponible, et des efforts communs doivent être faits par les scientifiques, les associations, les éditeurs, et les industriels, pour construire une ontologie chimique, unique, suffisamment générique et extensible, qui nous permette de transformer l'actuel système de documents et d'information en un système de représentation des connaissances.

Comme nous l'avons déjà indiqué, Il est important de rappeler que l'utilisation des langages de marquage n'est pas restreinte à la manipulation de l'information moléculaire. Ils sont utiles dans tous les aspects de l'informatique chimique, de la publication scientifique, de la transformation et traduction des données, de la construction des formats chimiques, de l'extraction et du traitement des données instrumentales, etc. La transformation des données chimiques actuelles en un système orienté vers la *connaissance* aura un effet considérable dans le traitement, la recherche, l'entretien et la réutilisation de l'information chimique future.

Conclusion

Dans ce chapitre nous avons traité en détail le lexique et la construction de la base de données des sous-structures, nécessaire à la construction des vecteurs descripteurs moléculaires. Des informations structurales et des propriétés ont été proprement encodées et structurées sous forme de fichiers XML. Ceci nous permettra plus tard d'extraire et de traiter cette information avec comme but d'effectuer des analyses de similarité et de diversité entre différents groupes de molécules.

Références

[Bebak1989] Bebak, H., Buse, C., Donner, W.T., Hoever, P., Jacob, H., Klaus, H., Pesch, J., Roemelt, J., Schilling, P., Woost, B., Zirz, C., *The Standard Molecular Data Format (SMD Format) as an integration tool in computer chemistry*, J. Chem. Inf. Comput. Sci. 29 (1989) 1-5.

[BinXML] Binary XML. Information disponible sur: http://www.expway.com/

[Bolev2001] Bolev, H., Decker, S., Sintek, M., *Tutorial on Knowledge Markup and Semantic Resources. IJCAI-01 (International Joint Conference on Artificial Intelligence)* Seattle, 6 Août 2001.

[Bremser1978] Bremser, W., *HOSE - a novel substructure code.* Anal. Chim. Acta, 103 (1978) 355 - 365.

[Carabedian1988] Carabedian, M., Dagane, I., Dubois, J.E. *Elucidation by Progressive Intersection of Ordered Structures from Carbon-13 Nuclear Magnetic Resonance.* Analytical Chemistry, 60 (1988) 2186-2192.

[CMLWeb] Chemical Markup Language (CML). Information disponible sur: http://www.xml-cml.org

[Dalby1992] Dalby, A., Nourse, J.G., Hounsell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A., Laufer, J, *Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited (MDL)*, J. Chem. Inf. Comput. Sci. 32 (1992) 244-255.

[DTDWeb] Document Type Definitions. Information disponible sur: http://www.xmlfiles.com/dtd/

[Dubois1980a] Dubois, J.E, Carabedian, M., Ancian, B. *Automatic structural elucidation by C-13 NMR - DARC-EPIOS method - Search for a discriminant chemical structure-displacement relationship.* Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences Serie C 290 (1980) 369-372.

[Dubois1980b] Dubois, J.E, Carabedian, M., Ancian, B. *Automatic structural elucidation by C-13*

NMR - DARC-EPIOS method - Description of progressive elucidation by ordered intersection of substructures. Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences Serie C 290 (1980) 383-386.

[Dubois1986] Dubois, J.E., Mercier, C., Panaye, A., *DARC topological system and computer aided design,* Acta Pharm. Jugosl., 36 (1986) 135-169.

[Dubois1999] Dubois, J.E., Doucet, J.P., Panaye, A., Fan, B.T., *DARC site toplogical correlations: ordered structural descriptors and property evaluation*. In Devillers, J. and Balaban, T. (Eds). Topological indices and related descriptors in QSAR and QSPR, Gordon and Breach Sciences Publishers, Amsterdam, 1999, pp. 613-673.

[Ertl2003] Ertl, P., *Chemoinformatics analysis of Organic Substituents: Identification of the most common substituents, calculation of substituent properties and automatic identification of Drug-like Bioisosteric Groups*, J. Chem. Inf. Comp. Sci. 43 (2003) 374-380.

[Gkoutos2001] Gkoutos, G.V., Murray-Rust, P., Rzepa, H.S. *The application of XML Languages for Integrating Molecular Resources*. Internet J. Chem. (2001) article 6.

[Harold2001] Elliot Rusty Harold, XML Bible, Wiley Eds., 2 edition, 2001.

[Murray-Rust1999] Murray-Rust, P., Rzepa, H.S., *Chemical Markup, XML and the Wold Wide Web. 1. Basic Principles.* J. Chem. Inf. Comput. Sci., 39 (1999) 928-942.

[Murray-Rust2001] Murray-Rust, P., Rzepa, H.S., *Chemical Markup, XML and the Wold Wide Web. 2. Information Objects and the CML-DOM*. J. Chem. Inf. Comput. Sci., 41 (2001) 1113-1123.

[Murray-Rust2002a] Murray-Rust, P., Rzepa, H.S., *Chemical Markup, XML and the Wold Wide Web. 3. Toward a signed Semantic Chemical Web of Trust.* J. Chem. Inf. Comput. Sci. 41 (2001) 1124-1130.

[Murray-Rust2002] Murray-Rust, P., Rzepa, H.S., *Markup Languages – How to Structure Chemistry-Related Documents*. Chemistry International, 4 (2002) 24-34.

[Murray-Rust2003] Murray-Rust, P., Rzepa, H.S., *Chemical Markup, XML and the Wold Wide Web. 4. CML Schema.* J. Chem. Inf. Comput. Sci. 43 (2003) 757-772.

[OWLWeb] Web Ontology language. Information disponible sur: http://www.w3.org/2004/OWL

[PerTableWeb] Information disponible sur: http://www.chem.qmw.ac.uk/iupac/AtWt/table.html

[RecomXMLWeb] Extended Markup Language (XML) 1.0, W3C Recommendation, 4 Février 2004. Information disponible sur: http://www.w3.org/TR/REC-xml

[Stobaugh1988] Stobaugh, R.E., *Chemical Abstract Service Chemical Registry System. 11. Substace-Related Statistics: Update and Additions*, J. Chem. Inf. Comp. Sci. 28 (1988) 180-187.

[TutorialXMLWeb] Tutorial en ligne de XML: http://www.w3schools.com/xml/xml_whatis.asp

[Weininger1988] Weininger, D., *SMILES (Simplified Molecular Input Line Entry System)*, J. Chem. Inf. Comput. Sci., 28 (1988) 31-36.

[Weininger1989] Weininger, D., Weininger, A., Weininger, J.L., *SMILES (Simplified Molecular Input Line Entry System)*, J. Chem. Inf. Comput. Sci., 29 (1989) 97-101. Information disponible sur: http://www.daylight.com/dayhtml/smiles

[Wiswesser1954] Wiswesser, W.J.A. (Ed.), A line-formula chemical notation, Crowell, New York, 1954.

[Xu2000] Xu, J., Stevenson, J., *Drug-like Index : A New approach to measure Drug like compounds and their Diversity*, J. Chem. Inf. Comput. Sci. 40 (2000) 1177-1187.

CHAPITRE III.

PROCESSUS DE COMPARAISON DES STRUCTURES MOLECULAIRES

III.1 Les recherches structurales
III.1.1 Algorithmes de superposition des graphes
III.1.2 Recherche de similarité pour des structures moléculaires
III.2 Reconnaissance des motifs structuraux et création des vecteurs descripteurs
III.2.1 Transformation des molécules et génération des vecteurs descripteurs
III.2.1.1 Reconnaissance des motifs structuraux
III.2.1.2 Génération des vecteurs-descripteurs

CHAPITRE III. PROCESSUS DE COMPARAISON

Dans ce chapitre nous expliquerons la manière dont nous avons effectué la comparaison des molécules et des fragments, ainsi que les critères qui contrôlent l'analyse.

III.1 Les recherches structurales

Quand on cherche des similitudes et des divergences entre les molécules, on peut effectuer plusieurs types de comparaisons: la recherche d'isomorphismes (sous-structures communes ou SSC), la recherche de la sous-structure maximum commune (SSMC) et la recherche de structures complètes (ou homomorphisme). Dans cette section, nous présentons des méthodes utilisées dans ces recherches. Certaines pourront être appliquées au développement de notre système.

Dans la partie suivante, nous parlerons de la recherche structurale SSC sur des graphes 2D qui permet de comparer deux structures « atome par atome ». Plusieurs algorithmes on été proposés dans la littérature pour simplifier et optimiser les possibilités de comparaison. Nous en présentons par la suite un échantillon.

III.1.1 Algorithmes de superposition des graphes

La recherche d'isomorphismes dite « recherche de sous-structures communes » ou C*ommon Substructure Search,* consiste à comparer une cible à un ensemble de sous-structures, avec comme but de trouver le nombre maximum de sous-structures communes à la cible. Généralement, ce type de recherche effectué dans des grandes bases de données est réalisé en deux étapes. D'abord, on présélectionne des structures candidates à l'aide d'un filtre pour ensuite effectuer des superpositions avec la cible en utilisant un algorithme adéquat. La qualité du filtre détermine la sélection des structures et limite le nombre de candidats retenus.

Nous énumérons ici brièvement quelques uns des algorithmes couramment utilisés dans la littérature pour effectuer la superposition d'un sous graphe 2D avec le graphe d'une structure candidate (recherche d'isomorphisme SSC). Les nœuds du graphe représentent ici les atomes, alors que les arêtes représentent les liaisons.

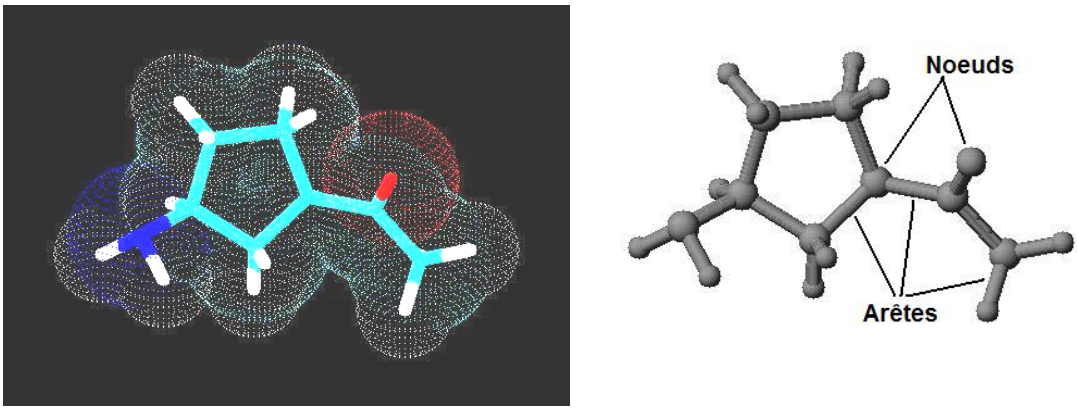


Figure III.1 Graphe d'une molécule : représentation mathématique simplifiée d'une structure chimique. Les graphes moléculaires sont couramment représentés en 2D. Ici la représentation est effectuée en 3D.

L'algorithme de Lesk [Lesk1979] est utilisé pour identifier les sous ensembles candidats possibles à l'isomorphisme avec une structure complexe. La congruence de ces sous ensembles avec la structure interrogée est testée. La première étape de l'algorithme consiste à identifier tous les atomes de la sous structure admissibles à la superposition avec chaque atome de la structure cible. Ensuite, tous les sous ensembles qui sont les candidats à la superposition avec les atomes de la structure interrogée sont générés. Plus on considère de propriétés pour les atomes, plus l'algorithme est efficace.

L'algorithme de Clique-détection [Barrow1976] sert à trouver des sous graphes maximaux complets d'une structure donnée (on entend par sous graphe maximal complet ou « clique » tout sous graphe complet dont tous les éléments ne sont pas contenus dans un autre sous graphe

complet). Ceci est réalisé en identifiant les parties communes à travers un marquage des nœuds et des arêtes des graphes 2D suivant le type d'atome et les distances inter atomiques. L'algorithme utilise la technique de *branch and bond* pour couper les branches qui ne peuvent pas conduire à une « clique ».

L'algorithme d'Ullmann [Ullmann1976] permet d'effectuer des recherches d'isomorphismes dans un ensemble de molécules. L'algorithme repose sur une recherche dans un arbre combinée avec l'élimination successive des nœuds afin d'augmenter l'efficacité.

Dans la première étape de l'algorithme, on génère les matrices M^0 construites en fonction des relations de correspondance entre l'ensemble des nœuds de deux graphes. Chaque élément de la matrice M^0 sera mis à 1 si les propriétés du $j^{ème}$ nœud du graphe test peuvent englober toutes les propriétés du $i^{ème}$ nœud du graphe cible, sinon il sera mis à 0. Dans la deuxième étape, on teste l'isomorphisme pour chaque matrice d'après une relation de superposition générant ainsi les matrices M^1. Ici tous les 1 sont changés par des zéros à l'exception d'un élément par rang (celui qui accomplit une superposition complète). A la fin de l'algorithme, un processus d'affinage est utilisé pour réduire le nombre de calculs nécessaires à la recherche d'un sous graphe isomorphe. Ainsi, pendant le parcours dans l'arbre, les nœuds des successeurs sont systématiquement éliminés.

L'algorithme de Sussenguth [Sussenguth1965] est utilisé pour rechercher des isomorphismes entre deux structures non-connexes.

La première étape consiste à générer des paires de sous ensembles de nœuds correspondants par référence à la sous-structure interrogée. Ces sous ensembles sont ensuite classés pour déterminer les correspondances nœud à nœud. Finalement, si la génération des sous ensembles est effectuée avec succès, la procédure de classification peut être lancée. Sinon, il faut voir si tous les nœuds de la structure cible ont été parcourus. Dans le cas où il n'y a plus de nœud, cela signifie que la structure cible possède moins de nœuds que la sous-structure interrogée et qu'en conséquence, il n'existe pas

d'isomorphisme. S'il reste des nœuds non parcourus, l'algorithme essaie de trouver de nouvelles possibilités. Cette dernière étape sera répétée jusqu'à ce que tous les nœuds soient parcourus.

L'algorithme de Figueras [Figueras1972] sert à effectuer des isomorphismes entre graphes. Son mécanisme d'action repose sur la théorie des ensembles et l'algèbre Booléenne.
Quand les propriétés des atomes sont codées et ordonnées, ces informations peuvent être analysées pour rejeter les codes non compatibles dans la recherche d'isomorphismes. Dans cet algorithme, la taille de la structure cible est réduite progressivement. Lorsque l'ensemble ne peut plus être réduit, le processus de comparaison s'arrête. La recherche inversée ou *back tracking* n'est pas utilisée dans cet algorithme, et par conséquent, son exécution est très rapide.

Les algorithmes faisant l'objet de modifications ces dernières années sont nombreux. Soit pour les adapter à des applications particulières, soit pour des raisons d'optimisation, les modifications et les combinaisons des procédures ne se comptent plus. Par exemple, les algorithmes pour la recherche d'isomorphismes 2D peuvent être appliqués avec quelques variantes aux structures chimiques en 3D. Ceci est possible en prenant en compte que les nœuds du graphe représentent toujours les atomes des structures ou des fragments moléculaires, et que les arêtes du graphe peuvent représenter les liaisons en 2D ou les distances réelles entre deux atomes en 3D.
Une adaptation de l'algorithme de « clique-détection » a été effectuée par Bron et Kerbosh [Bron1973] pour comparer efficacement des graphes chimiques et calculer des distributions de similarité et de diversité dans des librairies. Le même algorithme a été utilisé pour comparer des graphes 2D et pour évaluer d'autres paramètres structuraux comme la chiralité moléculaire et l'identification des degrés de liberté internes. Un autre algorithme alternatif repose sur l'algorithme de « Sussenguth ». Celui ci peut être utilisé pour faire des recherches SSC. Toutefois cette modification est valable seulement pour la recherche de petites sous-structures, car la place mémoire nécessaire est presque double à celle utilisée dans l'algorithme de Sussenguth.

Un autre algorithme souvent utilisé pour faire des recherches structurales mais en 3D, est l'algorithme de Crandell et Smith [Crandell1983]. Il utilise un processus itératif pour trouver la sous-structure commune maximale des sous structures 3D parmi un groupe de structures où toutes les sous structures communes ont une taille particulière. Le principe est d'ajouter peu à peu des atomes et d'éliminer ceux qui ne conviennent pas dans la recherche des candidats

III.1.2 Recherche de similarité pour des structures moléculaires

Les recherches d'homomorphismes et d'isomorphismes dans les bases de données moléculaires font partie des techniques de criblage virtuel parmi les plus populaires. Quand les informations contenues dans ces bases chimiques sont limitées ou incomplètes, il arrive souvent de ne pas trouver l'information relative à la structure désirée. Dans ce cas aucune réponse n'est obtenue, montrant ainsi les limitations des méthodes de recherche exacte de sous-structures. Les recherches de similarité étendent l'univers chimique des résultats en utilisant les informations des bases moléculaires pour trouver des structures « voisines » de la cible, tant du point de vue structural que des propriétés.

Dans l'introduction, nous avons déjà défini les concepts de similarité et de diversité mais d'une manière très générale, pour ensuite énumérer les éléments principaux d'une recherche de similarité dans un cadre moléculaire. Dans les sciences expérimentales, la similarité est mesurée selon les propriétés des objets. En mathématique, la similarité est classée en 5 catégories : analogie attributive, analogie fonctionnelle, analogie inductive, analogie proportionnelle et analogie structurale [Rouvray1990]. Il est important de noter que pas toutes les définitions de l'analogie mathématique sont applicables aux molécules et entités chimiques (atomes, liaisons, fragments, etc.).

- L'**analogie attributive** s'applique à des objets A et B qui ont respectivement les propriétés ou attributs *a* et *b*. Quand une propriété représente l'autre, ou plus généralement, s'il existe des correspondances entre les deux propriétés, ces deux propriétés sont analogues.
- L'**analogie fonctionnelle** s'applique à des objets A et B ayant une fonction commune ou possédant des composants jouant le même rôle.
- L'**analogie inductive** s'applique à une série d'objets, A, B, C, etc. Ces objets possèdent des propriétés communes, par exemple P et Q. Si les objets A et B possèdent également une autre propriété R, nous pouvons déduire, selon la logique inductive que l'objet C a probablement aussi cette propriété R.
- L'**analogie proportionnelle** s'applique aux propriétés de l'objet qui sont proportionnelles. Pour les objets A, B, C et D qui ont respectivement les propriétés *a, b, c* et *d*, l'analogie proportionnelle peut être exprimée avec la notation *a:b = c:d*, c'est-à-dire que « *a* est proportionnel à *b*, comme *c* est proportionnel à *d* ».
- L'**analogie structurale** s'applique à deux systèmes dont l'un est un modèle de l'autre. Les deux systèmes peuvent être décrits par le même ensemble d'égalités. Cette analogie s'appelle aussi « isomorphisme ».

La chimie a depuis longtemps recours à l'**analogie attributive**. Déjà en 1869, Dmitrii Mendeleïev proposait l'arrangement des éléments chimiques sous la forme d'une table périodique [MendeleïevWeb]. La base de son raisonnement était la similarité des propriétés, partagée par des groupes d'éléments.

Le concept d'analogie attributive est utilisé par les chimistes sous la forme du « principe de similarité des propriétés ». Comme nous l'avons indiqué dans le Chapitre I, ce principe établit que des molécules ayant une structure similaire ont plus de chances d'avoir des propriétés similaires par rapport à deux molécules choisis au hasard [Johnson1990, Walters1998, Martin2002].

Par exemple, les molécules i) 2-hydroxypropanamide et ii) 2-hydroxybutanamide (objets A et B), sont analogues car partageant une structure similaire. La molécule iii) (2E)-4,5-dimethylhex-2-ene, ne présente pas les mêmes analogies structurales que i et ii. La figure III.2 montre que pour les molécules i et ii, l'enthalpie standard de formation (Δ°f), est négative et proche de -440 KJ/mol, et que la polarisabilité est analogue (ces deux propriétés étant les attributs *a* et *b)*. La molécule iii ne présente pas de correspondances avec i et ii du point de vue des attributs étudiés.

Le comportement des molécules i et ii est connu en chimie comme « les séries homologues » : des atomes ou des molécules appartenant à la même « série » auront des propriétés voisines.

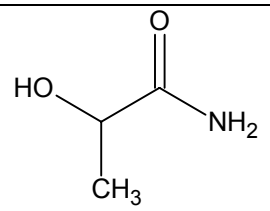
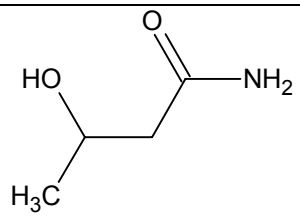
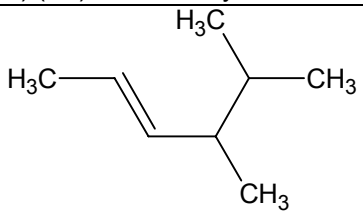
	Polarisabilité (1)	Δ°f (2)
i) 2-hydroxypropanamide 	$8.32 \pm 0.5\ 10^{-24}$ cm^3	- 421.8 kJ/mol
ii) 3-hydroxybutanamide 	$10.16 \pm 0.5\ 10^{-24}$ cm^3	- 457.33 kJ/mol
iii) (2*E*)-4,5-dimethylhex-2-ene 	$15.49 \pm 0.5\ 10^{-24}$ cm^3	-100.07 kJ/mol

Figure III.2. Analogie attributive entre paires de molécules et valeurs des propriétés physicochimiques. (1) Polarisabilité calculé avec ACDC/ChemSketch 5.12, (2) Enthalpie de formation calculé avec HyperChem 6.0 (méthode AM1).

Dernièrement, plusieurs auteurs ont souligné, à partir de résultats contradictoires, que le « principe de similarité des propriétés » devait être appliqué avec quelques précautions. La définition de similarité utilisée dans chaque situation doit être choisie soigneusement et adaptée aux besoins du calcul car il n'existe pas de règle absolue en termes de similarité pour le calcul des analogies moléculaires.

Des calculs effectués par Doucet *et al.* [Doucet1998] avec des composés liant le récepteur d'adénosine A1 (figure III.3), ont montré que malgré le fait que les molécules semblent voisines d'un point de vue structural, elles montrent des différences significatives, en particulier dans leur potentiel électrostatique moléculaire (MEP). Dans la figure III.4 (a) on observe que la meilleure correspondance entre les molécules A et B coïncide avec la prédiction structurale. Par contre en (b), la molécule C doit être tournée de 180° pour obtenir une bonne superposition des points MEP.

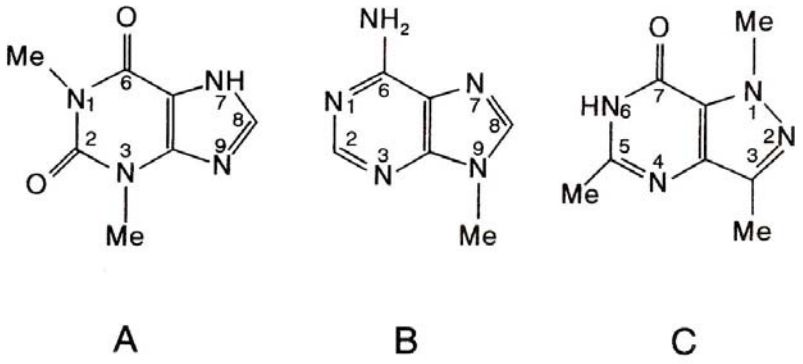


Figure III. 3 Molécules utilisés dans le test de comparaison. A : théophylline, B : adénosine et C : 5-(2-amino-4cholophenyl)-1,6-dihydro-1,3-imethyl-7H-pyrazolo(4,3-d) pyrimidin-7-one.

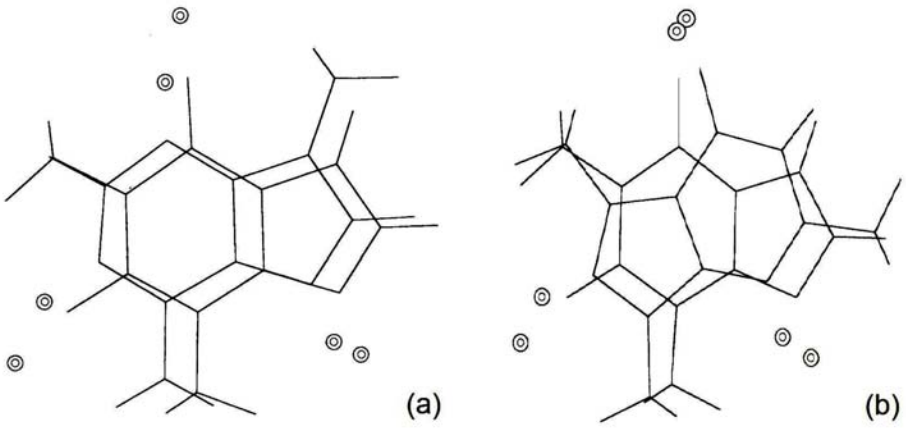


Figure III.4. Correspondance entre la théophylline et les molécules B (a) et C (b) considérant des propriétés électroniques (représentés par les points MEP) et les propriétés stériques. Le squelette atomique a été légèrement déplacé pour améliorer la visibilité de la superposition.

D'autres calculs menés par Gund *et al.* [Gund1980] et qui impliquent les anneaux ptéridine d'acide di-hydrofolique et de méthotrextate (figure III.5) ont été confirmés par Doucet *et al*.

A première vue, les deux molécules qui initialement présentent une structure 2D similaire semblent être de bons candidats pour se lier à la dihydrofolate réductase (DHFR). Mais une inspection visuelle de ses régions MEP (figure III.6) et l'évidence cristallographique confirment que dans la forme active, le méthotrextate a subi une rotation de 180°.

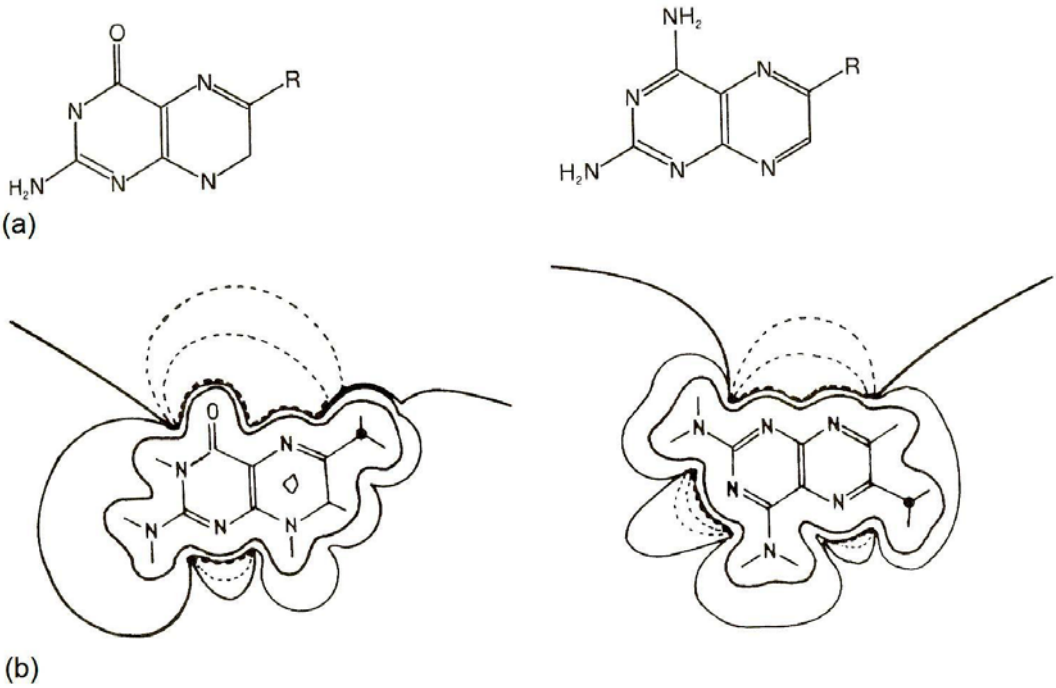


Figure III.5. (a) Molécules modèles d'acide di-hydrofolique et de méthotrextate. (b) Modèle MEP pour R = CH3. Les lignes pointillées correspondent aux régions MEP négatives.

Une étude assez récente des exceptions au « principe de similarité des propriétés » a été effectuée par Nikolova *et al.* [Nikolova2003]. Les auteurs affirment que l'usage de ce principe, basique pour l'étude et la prévision de l'activité chimique doit être fait attentivement. Plusieurs exemples sont donnés dans l'étude [BajorathWeb, TurinWeb, Meylan1999], parmi lesquels la figure III.6, qui met en évidence une liste de molécules choisies pour démontrer que des modifications sur la structure d'un composant peuvent produire des changements considérables des propriétés physicochimiques (et en conséquence de l'activité biologique). La formalisation du « principe de similarité de propriétés » à travers une métrique choisie permet donc de circonscrire l'application au problème étudié.

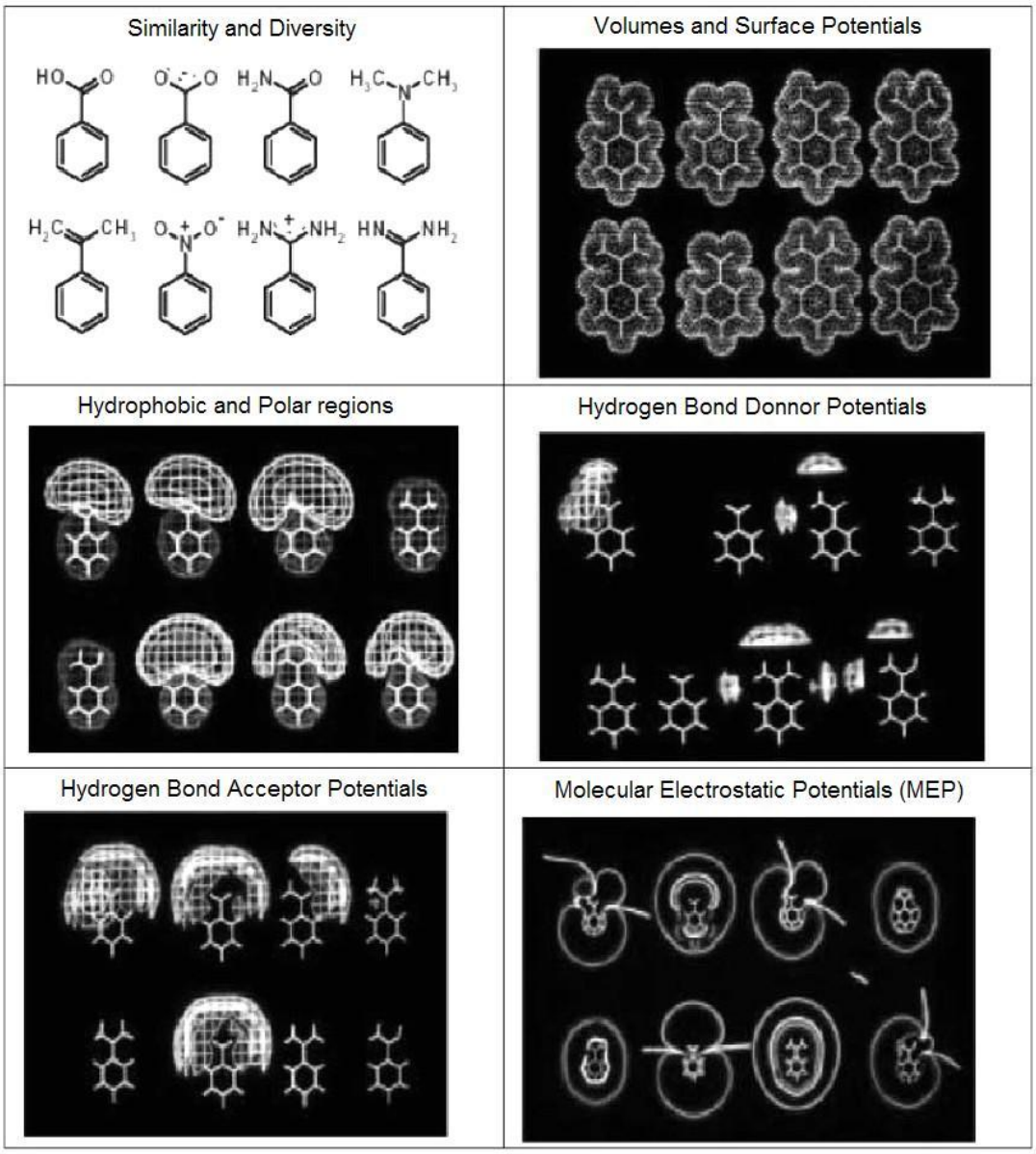


Figure III.6. Echantillon des composés structuralement similaires qui ont de grandes différences de volume, de potentiel de surface, de régions hydrophobiques et polaires, de potentiels d'accepteur ou de donneur de liaison hydrogène, et du potentiel électrostatique moléculaire (MEP) [Nikolova2003].

Le « principe de similarité des propriétés » étant prouvé non infaillible, les chimistes se sont tournés vers le concept d'**analogie inductive,** pour la recherche de similarités entre molécules. Ce concept appliqué en chimie permet de comparer directement plusieurs molécules avec des points de vue très divers (propriétés physicochimiques, activités, structures moléculaires, etc.). A partir d'une

métrique déterminée pour un groupe de molécules défini, on construit un modèle qui permet d'analyser des nouvelles molécules. Ceci étant un problème de régression (et d'extrapolation), on peut appliquer les méthodes d'apprentissage pour le résoudre. QSAR et QSPR sont en conséquence des applications des concepts inductifs.

Finalement, l'application de l'**analogie structurale** n'a été possible en chimie que depuis l'usage de modèles structuraux comme les graphes moléculaires (voir figure III .1). Ceux-ci ont permit de calculer non seulement des analogies, mais toutes les opérations mathématiques qui sont couramment appliquées aux graphes.

La construction des vecteurs moléculaires qui sont utilisés dans notre outil repose sur l'analogie structurale et l'analogie attributive. Cette comparaison exploite la différence (présence ou absence) de caractéristiques ou de propriétés déterminées ainsi que les différences structurales comme mesures de comparaison. Mais ce seront les mêmes définitions d'analogie mathématique qui nous permettront plus tard de gérer les vecteurs descripteurs générés par l'outil.

III.2 Reconnaissance des motifs structuraux et création des vecteurs descripteurs

Dans le chapitre II, nous avons expliqué la manière dont les bases de données qui composent notre outil ont été créées (section II.1) et comment elles sont structurées (section II.2). Nous rappelons donc que la FragDB et les QueryDB-TestDB sont les éléments clés de notre outil. L'une de ces bases permet d'extraire les fragments moléculaires qui nous intéressent, et l'autre fournit des molécules cibles ou tests à comparer. La création et le remplissage de l'index-XML (section II.2.2.1) sont aussi très importants, car cet index contient, sous une forme structurée, toutes les informations sur les fragments nécessaires au bon déroulement de la comparaison. Finalement, la création et le remplissage de l'indexResult-XML (section II.2.3.1) sont l'un des objectifs principaux de notre outil. Dans ce chapitre nous donnerons donc les détails de la transformation des molécules visant à la création des vecteurs descripteurs (voir figure II.7)

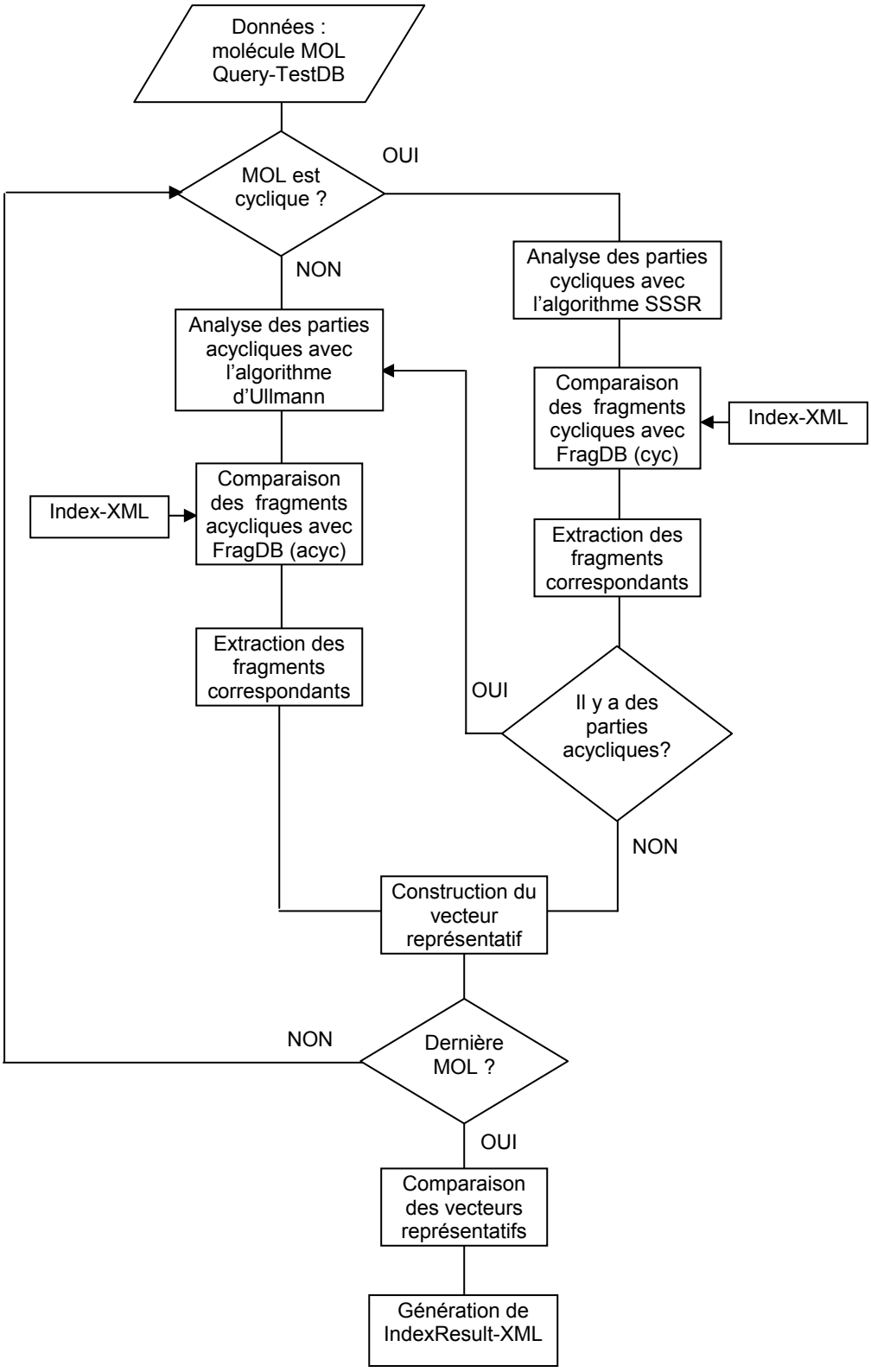


Figure III.7 Schéma général de la procédure de création des vecteurs moléculaires dans l'outil d'analyse de similarité et diversité des molécules proposé.

Dans un premier temps, la comparaison est de type structural. Elle est faite en utilisant l'algorithme d'Ullmann et un algorithme de type SSSR (*smallest set of smallest rings*), et elle vise à la création des vecteurs représentatifs (descripteurs). Dans un deuxième temps, la comparaison est faite entre ces vecteurs descripteurs en utilisant des indices, des coefficients ou des distances et elle vise au calcul de la mesure de similarité/diversité entre les deux molécules. Le schéma général de la procédure est donné dans la figure III.7.

III.2.1 Transformation des molécules et génération des vecteurs descripteurs

Pour effectuer la transformation des molécules en utilisant la FragDB, nous suivons les étapes suivantes :

- Reconnaissance des motifs structuraux des molécules contenues dans les molécules.
- Génération des vecteur-descripteurs pour l'ensemble des molécules analysées.

Par la suite, nous allons expliquer chaque étape en détail.

III.2.1.1 Reconnaissance des motifs structuraux

Pour faire une reconnaissance des motifs structuraux ou des fragments entre deux molécules, il est nécessaire d'effectuer une comparaison atome par atome. Pour comparer les atomes de la molécule cible avec ceux des fragments disponibles, nous avons utilisé des méthodologies différentes selon le type de molécule : acyclique ou cyclique.

a) Molécules acycliques : utilisation d'une méthodologie inspirée de l'algorithme d'Ullmann (déjà introduit dans la section III.1.1)

Avant d'expliquer l'algorithme pour la recherche des isomorphismes que nous avons appliqué dans notre outil, quelques éclaircissements sont nécessaires concernant la nature de la structure de données moléculaires et les conditions particulières des molécules cibles.

Comme nous l'avons indiqué auparavant dans la section II.2.3.1, des restrictions s'appliquent quant à la nature et au format des molécules des QueryDB-TestDB. Ainsi, il n'est pas nécessaire de représenter les hydrogènes liés aux atomes lourds car ces hydrogènes peuvent être générés automatiquement à partir des informations enregistrées pour les atomes lourds.

La structure de données des molécules construite à partir des informations stockées dans le fichier MOL est précisée dans le tableau suivant. Ces informations sont déterminées automatiquement à partir d'une modification du « module d'acquisition de formules structurales » qui avait déjà été mis au point au laboratoire ITODYS par Yao *et al.* pour le système de simulation spectrale IR et Raman « SIRS-SS » [Yao2001].

Propriétés	**Définitions**
Symbole d'élément	Numéro atomique des atomes selon la table périodique des éléments. Pour les atomes génériques voir le tableau II.2
Charge	Charge formelle, positive ou négative. -99 : charge incertaine ou non précisée.
Degré d'hybridation	SP : 1, =C=, −C≡ SP : 2, >C= SP : 3, >C<, SP : -99, Degré d'hybridation incertain ou non précisé.
Connectivité	Connectivité réelle. -99 : connectivité incertaine ou non précisée.
Valence	Valeur de la valence. -99 : valence incertaine ou non précisée.
Nombre d'H attachés	Somme exacte. -99 : somme incertaine ou non précisée.
Caractère isotopique	Valeur exacte. -99 : caractère isotopique incertain ou non précisé.
Caractère aromatique	Valeur binaire. 1 : caractère aromatique 0 : caractère non aromatique -99 : caractère aromatique incertain ou non précisé.

Tableau III.1 Caractéristiques des atomes et de leurs valeurs, selon la structure de données construite à partir des informations stockées dans les fichiers MOL.

L’algorithme pour la recherche des isomorphismes est ensuite décrit en utilisant un exemple. Dans la figure III.8 une structure cible et des fragments de la FragDB dont on cherche s’ils sont contenus dans la molécule cible sont représentés. Nous pouvons constater visuellement que seuls les deux premiers fragments de la base de données sont effectivement présents dans la molécule. L’algorithme d’Ullmann réalise automatiquement cette opération en comparant les caractéristiques des atomes. On donne les listes de caractéristiques des atomes, les matrices M créées pour comparer les atomes, et finalement, les matrices résultantes, une fois que l'algorithme d'Ullmann a détecté les fragments.

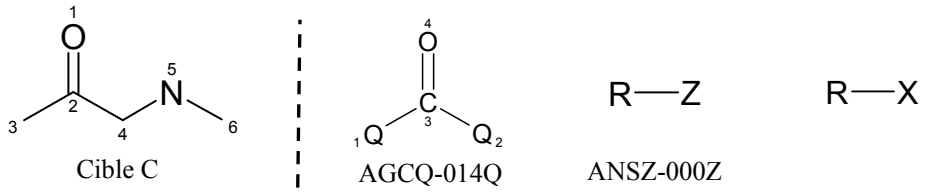


Figure III.8 Molécule cible C: 1-Methylamino-propan-2-one et 3 fragments de la FragDB. Pour le détail des atomes génériques Q, R, Z, X voir le tableau II.1.

La figure III.9 montre à travers un cas d’étude (l’analyse atome par atome de la 1-Methylamino-propan-2-one et des trois fragments contenus sur la FragDB) les étapes de l’analyse d’une molécule par notre outil.

La première étape est donc l’extraction des propriétés. Celles-ci seront calculées à partir de la lecture du fichier source (MOL) et des informations de connectivité et de type d’atome. Ceci est la raison pour laquelle on ne peut pas lire d’autres formats de fichier que le format MOL, au risque que la lecture des informations soit erronée. Des outils qui transforment les nombreux formats moléculaires existant (par exemple : Babel) résolvent ce problème.

(a) **Propriétés des atomes**	**Cible C**						**AGCQ-014Q**				**ANSZ-000Z**		**ANSX-000X**	
	O_1	**C_2**	**C_3**	**C_4**	**N_5**	**C_6**	**Q_1**	**Q_2**	**C_3**	**O_4**	**R**	**Z**	**R**	**X**
Symbole d'élément	8	6	6	6	7	6	146	146	6	8	138	140	138	142
Charge	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Degré d'hybridation	2	2	3	3	3	3	3	3	2	2	3	3	3	3
Connectivité	1	3	1	2	2	1	1	3	1	1	1	1	1	1
Valence	4	4	4	4	3	4	-99	-99	4	4	-99	-99	-99	-99
Σ des H attachés	0	0	3	2	1	3	-99	-99	0	0	-99	-99	-99	-99
Caractère isotopique	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Caractère aromatique	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Algorithme d'Ullmann

(b) **Matrice Ullmann pour cible C**	**AGCQ-014Q**				**ANSZ-000Z**		**ANSX-000X**	
	Q_1	**Q_2**	**C_3**	**O_4**	**R**	**Z**	**R**	**X**
O_1	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**
C_2	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**
C_3	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
C_4	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**
N_5	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**
C_6	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**

Détection des fragments

(c) **Matrice Ullmann pour cible C**	**AGCQ-014Q**				**ANSZ-000Z**		**ANSX-000X**	
	Q_1	**Q_2**	**C_3**	**O_4**	**R**	**Z**	**R**	**X**
O_1	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**
C_2	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**
C_3	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
C_4	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**
N_5	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**
C_6	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**

Figure III.9 (a) Propriétés des atomes des molécules de la figure III.8. Pour faciliter la lecture, les atomes sont numérotés de la même manière que dans le fichier MOL. (b) Table des correspondances atome par atome après passage de l'algorithme d'Ullmann (c) Avec les informations de la molécule cible on regroupe les fragments détectés.

Les propriétés qui ont été extraites font partie des caractéristiques minimales nécessaires à l'identification d'un atome, à savoir, le type d'atome, l'hybridation, la charge, la connectivité, la valence, les H attachés, les isotopes et le caractère aromatique. Nous avons essayé d'extraire un ensemble minimal et le moins redondant possible des caractéristiques envisageables. Les valeurs assignées correspondent aux valeurs réelles extraites ou calculées du fichier MOL. Si la valeur n'est pas disponible (incertaine ou non précisée), la valeur -99 est assignée.

Une fois toute l'information extraite, on l'ordonne sur une structure de données vide, interne à l'outil. Cette structure de données sera la base de tous les traitements effectués dans le futur sur la (les) molécule(s) cible(s) et elle permettra d'identifier les fragments de la FragDB.

La seconde étape est celle de la comparaison atome par atome. Ceci sera effectué en examinant les propriétés explicitées dans le tableau III.1 et en appliquant des filtres de plus en plus sélectifs, pour optimiser la comparaison. Par exemple, si l'atome C_3 est comparé à l'atome O_1, un premier test sur le type d'atome sera négatif et il ne sera plus nécessaire de continuer la comparaison entre les deux atomes. Par contre, une comparaison entre l'atome Q_1 et N_5 passera les premiers tests mais échouera au test de connectivité. Plus l'atome est générique, plus il faut aller loin dans le test pour être sur que les deux atomes soient identiques ou au moins comparables.

La troisième étape est la création d'une matrice vide de dimensions adaptées, dans laquelle on enregistrera les informations correspondantes aux comparaisons exactes et floues effectuées entre les atomes. Des doublons feront leur apparition, notamment au moment de comparer des fragments symétriques : par exemple, Q_1 et Q_2 seront reconnus tour à tour avec C_3 et C_4, mais une seule combinaison sera possible.

La quatrième étape consiste à appliquer notre modification de l'algorithme d'Ullmann sur l'ensemble des valeurs de la matrice pour écarter les faux doublons et trouver les isomorphismes.

b) Molécules cycliques : utilisation de l'algorithme SSSR puis de l'algorithme d'Ullmann.

L'algorithme pour la recherche des isomorphismes dans le cas des molécules cycliques comprend

deux parties principales : l'extraction des cycles moléculaires à travers l'usage d'un algorithme SSSR (*Smallest Set of Smallest Rings*) et le traitement des fragments cycliques et acycliques restants avec notre adaptation de l'algorithme d'Ullmann.

L'algorithme SSSR utilisé par notre outil correspond au « Ring perception SSSR » de B.T. Fan et al, [Fan1993, Petitjean2000]. Cet algorithme extrait le groupe minimal de cycles les plus petits pour une molécule déterminée. L'entrée de l'outil consiste en une molécule au format .MOL. La sortie est un fichier .CYC qui contient, en format texte, la liste des atomes appartenant aux cycles. Pour le détail des étapes de fonctionnement de l'algorithme, on peut consulter les références données plus haut.

Les systèmes cycliques, avec leur premier niveau d'atomes voisins, sont alors reconstruits et regroupés dans des molécules cycliques « temporaires » qui pourront être ensuite comparés avec la base des fragments cycliques (figure III.10). L'algorithme de reconnaissance des atomes voisins est détaillé plus bas.

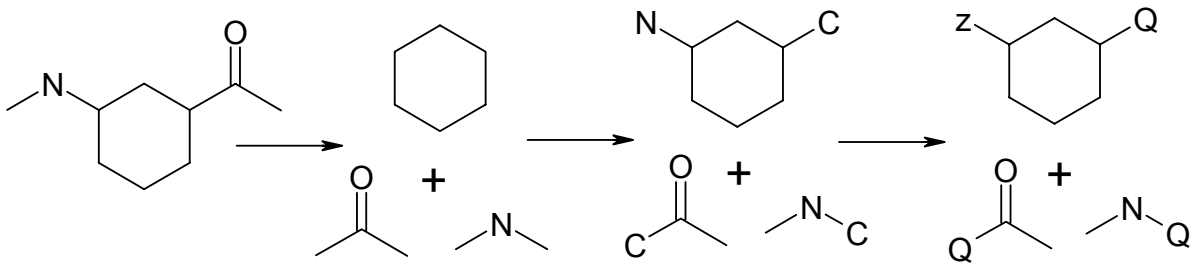


Figure III.10 Mécanisme de coupure d'une molécule : molécule entière, fragments obtenus, addition des atomes voisins et modifications subséquentes pour augmenter les correspondances.

Une fois les systèmes cycliques détectés et extraits de la liste d'atomes de la molécule cible, on effectue une reconstruction de la connectivité et des propriétés des atomes restants. Pour cela, on a utilisé le même algorithme de reconnaissance des atomes voisins déjà utilisé pour trouver les atomes voisins des cycles. Dans un premier temps, l'algorithme traite les voisins immédiats des atomes non-cycliques déjà identifiés. Ensuite, il analyse si les atomes appartiennent aux cycles ou non. Finalement, les structures de données des atomes choisis et de ses voisins sélectionnés sont

remplies. De cette manière, on transforme les atomes isolés n'appartenant pas à des cycles en nouvelles molécules à part entière. Ces molécules acycliques « temporaires » serviront de base à des futures comparaisons. La figure III.11 nous montre schématiquement les étapes de l'algorithme.

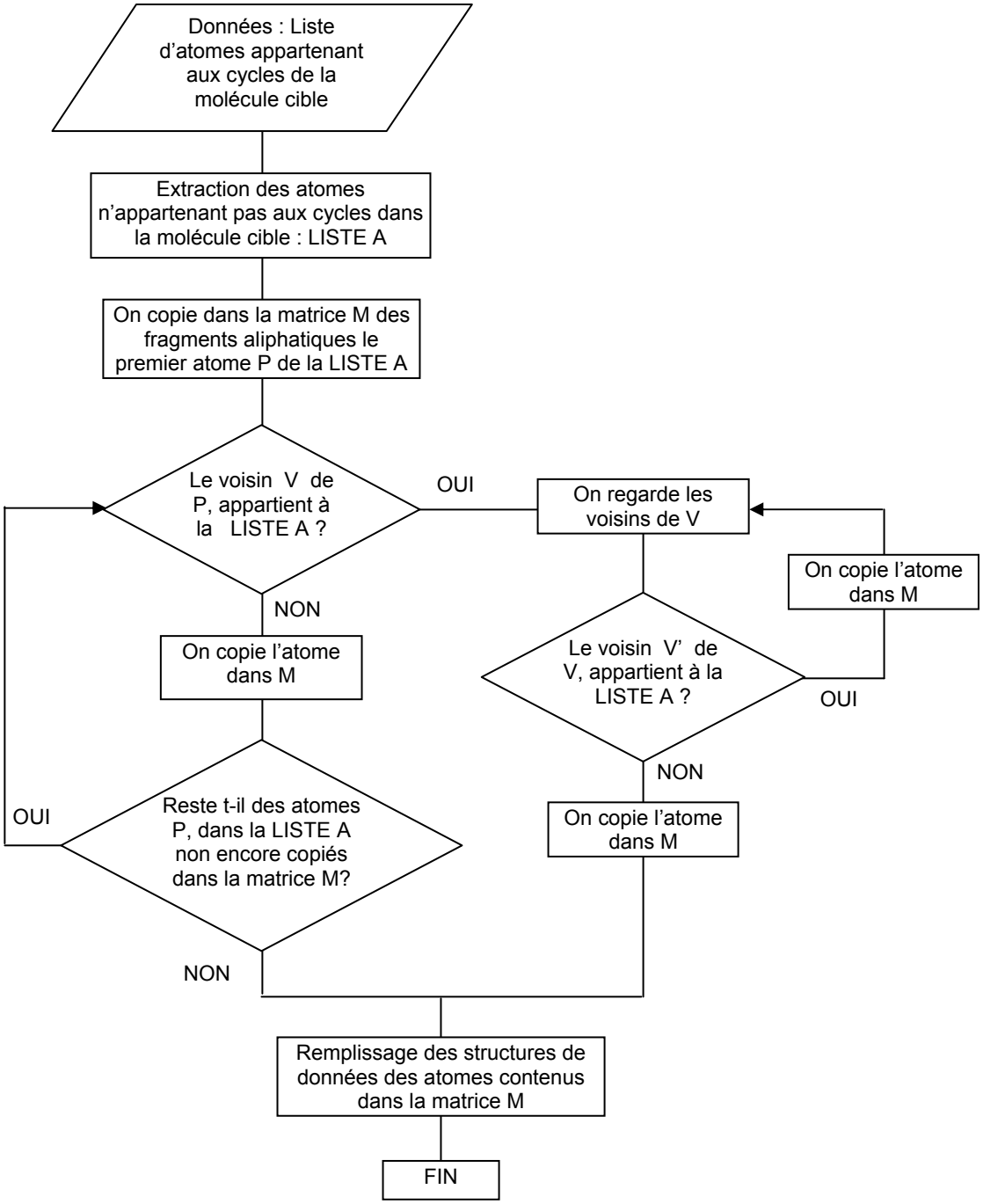

Données : Liste d'atomes appartenant aux cycles de la molécule cible
Extraction des atomes n'appartenant pas aux cycles dans la molécule cible : LISTE A
On copie dans la matrice M des fragments aliphatiques le premier atome P de la LISTE A
Le voisin V de P, appartient à la LISTE A ?
OUI
On regarde les voisins de V
On copie l'atome dans M
NON
On copie l'atome dans M
Le voisin V' de V, appartient à la LISTE A ?
OUI
NON
On copie l'atome dans M
OUI
Reste t-il des atomes P, dans la LISTE A non encore copiés dans la matrice M?
NON
Remplissage des structures de données des atomes contenus dans la matrice M
FIN

Figure III.11 Algorithme de reconstruction des fragments aliphatiques à partir des atomes isolés n'appartenant pas aux cycles de la molécule cible.

Une fois que la molécule cible a été découpée, on obtient une série de molécules fragmentaires (temporaires) cycliques et acycliques prêtes à l'analyse avec notre adaptation de l'algorithme d'Ullmann. Pour les parties cycliques de la molécule cible, on compare la partie cyclique de la base de fragments, et les sous-structures acycliques sont comparées à leur tour avec les fragments acycliques de la molécule analysée.

L'algorithme pour la recherche des isomorphismes entre la molécule cible et la FragDB a déjà été décrit dans la section III.2.1.1a. Maintenant, au travers d'un autre exemple nous allons illustrer le processus complet d'analyse d'une molécule cyclique. Ainsi, dans la figure III.12, une structure cible et des fragments de la FragDB dont on veut savoir s'ils sont contenus dans la molécule cible sont représentées. L'algorithme d'Ullmann détecte automatiquement les fragments de la molécule qui correspondent à ceux de la FragDB en comparant les caractéristiques des atomes. D'abord la molécule cible est décomposée en parties cycliques et non cycliques (voir figure III.13) selon l'algorithme montré sur la figure III.7.

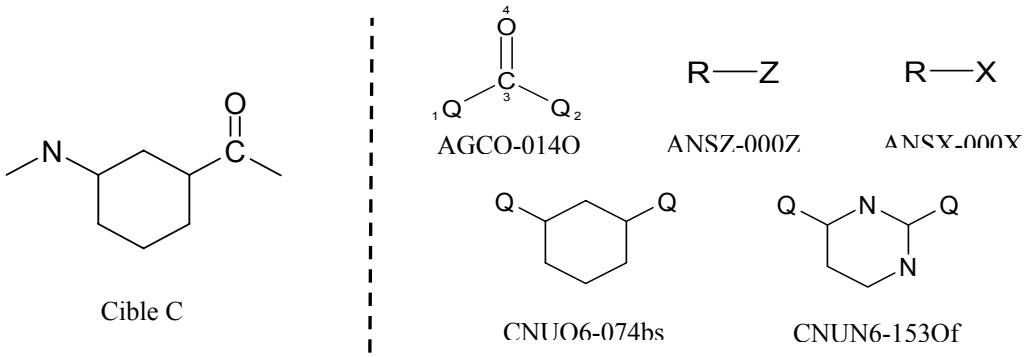


Figure III.12 Molécule cible C: 1-(3-Methylamino-cyclohexyl)-propanone et 5 fragments de la FragDB. Pour le détail des atomes génériques Q, R, Z, X voir le tableau II.1.

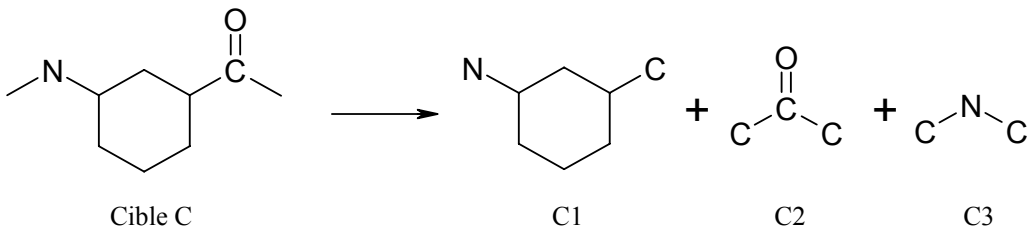


Figure III.13 Décomposition de la cible après application des algorithmes (SSSR + figure III.11)

Par comparaison des listes de caractéristiques des atomes (voir figure III.9a) on créé les matrices M qui serviront plus tard à construire les matrices résultant de l'application de l'algorithme d'Ullmann. La figure III.14 montre les deux dernières étapes du processus pour les parties de nature cyclique de la molécule cible C. La construction des matrices pour les parties non cycliques de la cible C est montrée dans la figure III.15.

(a) **Matrice Ullmann. Partie cyclique de la Cible C**		**CNUQ6-074bs**								**CNUN6-153Qf**							
		Q_1	Q_2	C_3	C_4	C_5	C_6	C_7	C_8	Q_1	Q_2	C_3	C_4	C_5	N_6	C_7	N_8
C1	C_1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
	C_2	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0
	C_3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	C_4	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0
	C_5	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
	C_6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	C_7	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
	N_{10}	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0

Détection des fragments

(b) **Matrice Ullmann. Partie cyclique de la Cible C**		**CNUQ6-074bs**								**CNUN6-153Qf**							
		Q_1	Q_2	C_3	C_4	C_5	C_6	C_7	C_8	Q_1	Q_2	C_3	C_4	C_5	N_6	C_7	N_8
C1	C_1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
	C_2	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0
	C_3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	C_4	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0
	C_5	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
	C_6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	C_7	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
	N_{10}	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0

Figure III.14 (a) Matrice après passage de l'algorithme d'Ullmann: comparaison atome par atome entre les parties cycliques de la cible C et les fragments cycliques de la figure III.12 (b) Matrice où l'on a regroupé les fragments détectés. En pointillé, un doublon pour le même fragment. La modification de l'algorithme d'Ullmann qui a été implémentée, identifie et élimine automatiquement tous les doublons des fragments détectés.

(a) **Matrice Ullmann. Parties acycliques de la Cible C**		**AGCQ-014Q**				**ANSZ-000Z**		**ANSX-000X**	
		Q_1	**Q_2**	**C_3**	**O_4**	**R**	**Z**	**R**	**X**
C2	**O_9**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**
	C_7	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**
	C_5	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
	C_8	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**
C3	**N_{10}**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**
	C_1	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**
	C_{11}	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**

Détection des fragments

(b) **Matrice Ullmann. Parties acycliques de la Cible C**		**AGCQ-014Q**				**ANSZ-000Z**		**ANSX-000X**	
		Q_1	**Q_2**	**C_3**	**O_4**	**R**	**Z**	**R**	**X**
C2	**O_9**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**
	C_7	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**
	C_5	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
	C_8	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**
C3	**N_{10}**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**
	C_1	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**
	C_{11}	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**

Figure III.15 (a) Matrice après passage de l'algorithme d'Ullmann: comparaison atome par atome entre les parties acycliques de la cible C et les fragments acycliques de la figure III.12 (b) Matrice où l'on a regroupé les fragments détectés. Les atomes sont numérotés de la même manière que dans le fichier MOL.

Les figures III.14 et III.15 illustrent, à travers l'analyse de 1-(3-Methylamino-cyclohexyl)-éthanone et de ses parties composantes C1, C2 et C3, les étapes dans l'analyse d'une molécule cyclique par notre outil.

Tout au long de cette section, nous avons expliqué la manière comment sont analysés les différent types de molécules cible, cycliques ou non cycliques, en utilisant les algorithmes modifiés

d'Ullmann et l'algorithme de SSSR pour obtenir une liste de fragments propres à chaque molécule et qui servira de descripteur moléculaire dans la section suivante.

III.2.1.2 Génération des vecteurs-descripteurs

Les matrices de résultats, après l'utilisation de l'algorithme d'Ullmann, permettent d'identifier les fragments contenus dans la FragBD présents dans la molécule analysée. Cette opération est effectuée très rapidement à cause de la simplicité de l'algorithme et des principes de la comparaison atome - atome. On extrait les données des fragments détectés à partir des résultats de la matrice d'Ullmann et on construit un vecteur résultat qui contient toute l'information concernant les fragments détectés et la molécule analysée.

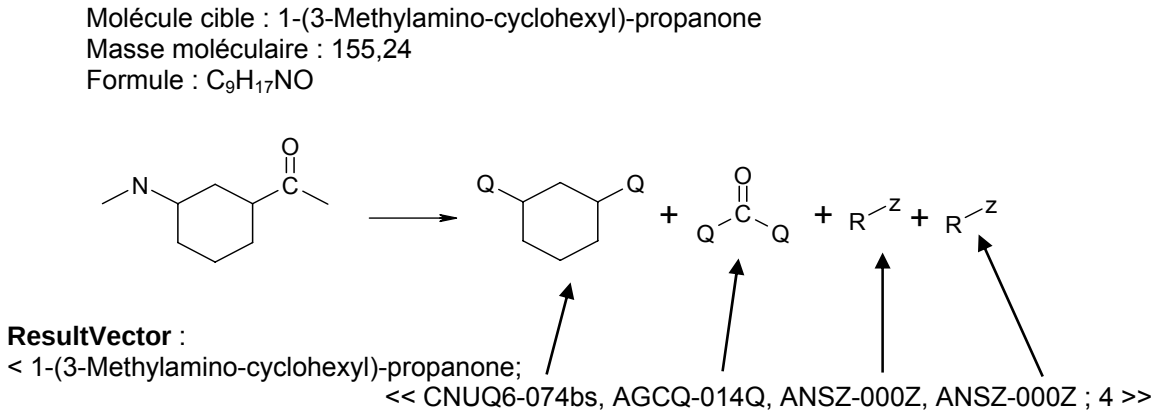


Figure III.16 Construction du vecteur résultat de 1-(3-Methylamino-cyclohexyl)-propanone à partir des informations extraites des matrices Ullmann.

Le vecteur descripteur (ou vecteur résultat) respecte une structure déterminée, conçue pour contenir les informations correspondant aux fragments, ainsi que leurs informations attachées (nom de fichier, propriétés physico-chimiques, etc.) Ce vecteur est donc structuré, comme montré dans la figure III.17. Le vecteur résultat ainsi obtenu sera dorénavant utilisé comme le descripteur structural pour la molécule analysée. On pourra alors effectuer une comparaison entre deux molécules en

utilisant leurs descripteurs respectifs. De plus, grâce au fait que le vecteur résultat contient des informations hétérogènes (données structurales et de propriétés physicochimiques), on peut effectuer des comparaisons moléculaires selon des critères très divers. Les différents critères de comparaison entre les molécules de la Query-TestDB sont le sujet du chapitre 4.

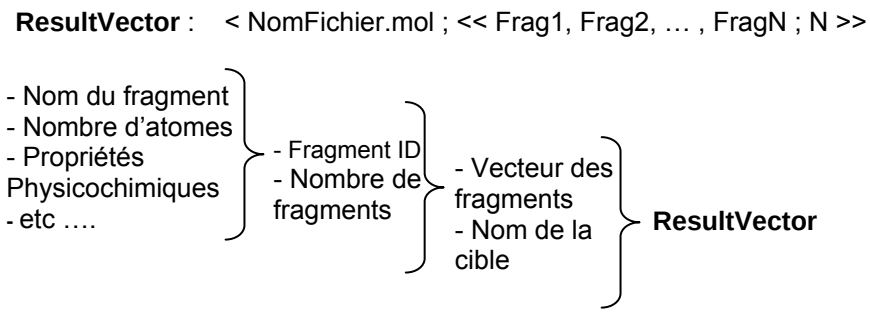


Figure III.17 Structuration du vecteur résultat.

Conclusion

Tout au long de ce chapitre nous avons expliqué la stratégie d'analyse moléculaire visant à construire des descripteurs structuraux. Ces descripteurs vont ensuite être utilisés pour effectuer des comparaisons entre molécules selon des critères structuraux ou liés aux propriétés.

Références

[BajorathWeb] Bajorath, J., *Virtual Screening in drug discovery: Methods, expectations and reality*. Information disponible sur: http://www.currentdrugdiscovery.com

[Barrow1976] Barrow, H.G. et Burstall, R.M., *Subgraph isomorphism, matching relational structures and maximal cliques,* Inf. Proc. Lett., 4 (1976) 83-84.

[Bron1973] Bron, C., Kerbosh, J., *Finding all cliques of an undirected graph*, Commun. ACM, 16 (1973) 575–577. Disponible sur l'URL: http://www.nap.edu/readingroom/books/mctcc/index.html

[Crandell1983] Crandell, C.W., Smith, D.H., *Computer-Assisted Examination of Compounds for Common Three-Dimensional Substructures,* J. Chem. Inf. Comput. Sci., 23 (1983) 186-197.

[Doucet1998] Doucet, J.P., Panaye, A., *3D Structural Information: form property prediction to substructure recognition with neural networks,* SAR and QSAR Envirom. Res., 8 (1998) 249-272.

[Fan1993] Fan, B.T., Panaye, A., Doucet, J.P., Barbu, A., *Ring perception. A new algorithm for directly finding the smallest set of smallest rings from a connection table.* Journal of Chemical Information and Computer Sciences 33 (1993) 657-662.

[Figueras1972] Figueras, J., *Substructure search by set reduction* J. Chem. Doc. 12 (1972) 237-244.

[GasteigerWEB] Disponible sur: http://www2.chemie.uni-erlangen.de/services/petra/smiles.phtml

[Gund1980] Gund, P., Andose, J.D., Rhodes, J.B., Smith G.M., *Three-Dimensional Molecular Modeling and Drug Design*, Science, 208 (1980) 1425-1431.

[Johnson1990] Johnson, A.M., Maggiora, G.M. (Eds.) Concepts and Applications of Molecular Similarity, John Willey & Sons, New York, Inc. 1990.

[Lesk1979] Lesk, A.M., *Detection of 3D patterns of atoms in chemical structures,* Comm ACM, 22 (1979) 219-224.

[Martin2002] Martin, Y.C., Kofron, J.L., Traphagen, L.M. *Do structurally similar molecules have similar biological activity?,* J. Med. Chem., 45 (2002) 4350-4358.

[MendeleïevWeb] Information disponible sur l'URL: http://pearl1.lanl.gov/periodic/mendeleev.htm

[Meylan1999] Meylan, W.M., Howard, P.H., Boethling, R.S., Aronson, D., Printup, H. et Gouchi, S., *Improved methods for estimating bioconcentration/ bioaccumulation factor from Octanol/Water partition coefficient,* Environ. Toxicol. Chem., 18 (1999) 664-672.

[Nikolova2003] Nikolova, N. et Jaworska, J., *Approaches to Measure Chemical Similarity - a Review,* QSAR Comb. Sci., 22 (2003) 1006-1026.

[Petitjean2000] Petitjean M., Fan B.T., Panaye A., Doucet J.P., *Ring perception: proof of a formula calculating the number of the smallest rings in connected graphs*, J. Chem. Inf. Comput. Sci., 40 (2000) 1015-1017.

[Rouvray1990] Rouvray, D.H., *The evolution of the concept of molecular similarity*. In Johnson, M.A. and Maggiora, G.M. (Eds.) Concepts and Applications of Molecular Similarity, John Willey & Sons, New York, Inc. 1990. pp. 15-42.

[Sussenguth1965] Sussenguth, E.H., *A Graph-Theoretic Algorithm for Matching Chemical Structures*, J. Chem. Doc. 5 (1965) 36-43.

[TurinWeb] Turin, L. et Fumiko, Y., *Structure-odor relations: a modern perspective*. Disponible sur l'URL: http://www.flexitral/research/review_final.pdf

[Ullmann1976] Ullmann, J.R., *An Algorithm for Subgraph Isomorphism,* J. ACM., 23 (1976) 31-42.

[Walters1998] Walters, W.P., Stahl, M.T. et Murcko, M.A. *Virtual Screening - An Overview*, Drug Discovery Today, 3 (1998) 160-178.

[Yao2001] Yao, J., Fan, B.T., Doucet, J.P., Panaye, A., Yuan, S. and Li, J., *SIRSS-SS: A system for Simulating IR/Raman Spectra. 1. Substructure/Subspectrum Correlation,* J. Chem. Inf. Comput. Sci., 41 (2001) 1046-1052.

CHAPITRE IV.

MESURES DE SIMILARITE MOLECULAIRES

IV.1 Coefficients et distances
IV.2 Comparaisons intermoléculaires
IV.2.1 Analyses de Similarité
IV.2.2 Calcul de la précision et du rappel « *recall* »
IV.3 Les différents niveaux de comparaison
IV.3.1 Comparaison exclusivement structurale
IV.3.2 Comparaison reposant sur la structure et les propriétés des molécules

CHAPITRE IV. MESURE DE SIMILARITE

Dans un contexte chimique les coefficients, les indices et les distances donnent une mesure quantitative du niveau de ressemblance entre deux modèles moléculaires. De très nombreuses représentations existent aujourd'hui pour exprimer la **comparaison** entre deux descripteurs moléculaires ou pour établir une **distance** entre deux objets dans un espace déterminé. Dans ce chapitre nous expliquerons le choix et l'usage de certaines de ces mesures pour le calcul de la similarité et de la diversité moléculaire.

IV.1 Coefficients et distances

Différents types de coefficient de similitude ont été décrits dans la littérature mais la plupart d'entre eux peuvent être regroupés en trois grandes classes : les mesures de distance, les coefficients d'association et les coefficients de corrélation. D'autres classifications ont été rapportées dans la littérature, notamment la classification en coefficients de corrélation, mesures probabilistes, associatives et de distances [Holliday2002], et la différenciation entre les indices de similarité symétriques et asymétriques [SimWeb]. Finalement, en fonction des données utilisées, on peut aussi les classer comme indices binaires et quantitatifs.

Les **mesures de distance** quantifient le degré de différence entre deux objets et ont été intensivement employées dans beaucoup d'applications des statistiques multi variées (particulièrement dans des cas où des variables à valeurs continues sont utilisées), en raison probablement de l'interprétation géométrique simple qui est attachée à bon nombre d'entre elles (par exemple, la distance euclidienne). Avec les mesures de distance, plus le degré de similitude entre deux objets est grand, plus la valeur du coefficient (de leur distance) est petite (et vice versa).

Les **coefficients d'association**, sont employés le plus généralement avec des données binaires (variables dénotant la présence ou l'absence des descripteurs dans un objet). Ils sont souvent

normalisés pour se situer dans un intervalle compris entre zéro (aucune similitude du tout, différence maximale) et l'unité (ensembles identiques de descripteurs). Cela dit, les coefficients d'association peuvent être employés avec des données non-binaires. Dans ce cas, d'autres gammes de valeurs peuvent s'appliquer ou de nouvelles constantes de normalisation être utilisées.

D'autre part, les **coefficients de corrélation** mesurent le degré de corrélation entre les ensembles de valeurs caractérisant une paire d'objets. D'autres utilisations plus conventionnelles incluent les analyses multi variées où l'on recherche les rapports entre des paires de variables.

Parmi le grand nombre de coefficients et de distances de similarité définis, beaucoup sont inter-dépendants. Il arrive que certains coefficients puissent être obtenus par des approches différentes. D'autres ont des comportements similaires en fonction des données employées (binaires, réels, etc). On assigne donc le terme « monotone » [Willet1987] aux coefficients ou distances de similarité qui montrent des résultats analytiques équivalents et donc un ordre de classement (*ranking*) identique, pour un groupe de molécules donné. Formellement, deux mesures de similarité S1 et S2 sont monotones lorsque, pour tout couple de molécules i, j on a:

$$(S1(i)-S1(j)) \times (S2(i)-S2(j)) \geq 0$$

Autrement dit, les quantités S1(i)-S1(j) et S2(i)-S2(j) sont positives ensemble ou négatives ensemble. On vérifie que la relation de monotonie est réflexive (S est monotone avec S), symétrique (S1 monotone avec S2 implique S2 monotone avec S1), et transitive (S1,S2 monotones, et S2,S3 monotones => S1,S3 monotones). Même s'il est rare de trouver deux coefficients 100% monotones, il suffit d'un haut degré de corrélation entre les résultats des deux coefficients dans un groupe de molécules, pour reconsidérer leur usage ou les utiliser d'une manière complémentaire. Tout au

contraire, les coefficients ou distances qui affichent une très faible corrélation expriment, par leurs résultats, des caractéristiques différentes des molécules qui sont comparées.

De nombreux travaux font état des mesures les plus communément utilisées [Willett1987, Willett1998]. La pertinence des différents coefficients de similarité a été également le sujet de nombreux travaux [Pearlman1999, Willet1986, Holliday2002]. Dans le tableau suivant, nous résumons quelques unes des mesures répertoriées dans la bibliographie.

Type de Coefficient	Nom	Expression
Coefficient associatif	Cosinus	$S_C = \frac{c}{\sqrt{ab}}$
Coefficient associatif	Forbes	$S_F = \frac{cn}{ab}$
Coefficient associatif	Russell-Rao	$S_R = \frac{c}{n}$
Coefficient associatif	Simpson	$S_{SI} = \frac{c}{\min(a,b)}$
Coefficient associatif	Tanimoto	$S_T = \frac{c}{a+b-c}$
Coefficients de corrélation	Yule	$S_Y = \frac{nc-ab}{cd+(a-c)(b-c)}$
Coefficients de corrélation	Dennis	$S_D = \frac{nc-ab}{\sqrt{nab}}$
Coefficients de corrélation	Pearson	$S_P = \frac{nc-ab}{\sqrt{ab(n-b)(n-a)}}$
Distance	Squared Euclidean	$S_E = \frac{a+b-2c}{n}$

Tableau IV.1 Exemples de quelques coefficients de Similarité/Diversité

Pour évaluer la similarité entre deux molécules avec les formules indiquées dans le Tableau IV.1 il faut définir les variables suivantes de manière générale :

a : représente le nombre d'entités de la première molécule.
b : représente le nombre d'entités de la deuxième molécule.
n : est le nombre total d'entités (dimension/longueur du vecteur descripteur)
c : est le nombre d'entités communes aux deux molécules
d : est le nombre d'entités non communes entre les deux molécules.

Par « entité», on n'entend pas seulement des propriétés physicochimiques, mais aussi des propriétés structurales ou de forme, par exemple des fragments... La coïncidence entre les éléments de ces « propriétés » sera interprétée comme une partie importante dans les mesures de similarité.

On note également que la liste des indices et distances présentée au tableau IV.1, a été restreinte à ceux qui peuvent être calculés de manière suffisamment efficace pour être utilisés dans le traitement de grandes bases de molécules, la motivation initiale de ce travail étant l'application de l'outil dans le criblage virtuel de grandes bases de données.

On remarque assez souvent dans la bibliographie [Holliday2003, Willett1998, Whittle2004] que l'indice de Tanimoto est préféré, même si quelques avis défavorables ont été formulés. [Dixon1999], et [Lajiness1997] lui reprochent notamment des résultats biaisés pour les petites molécules quand des analyses de diversité sont effectuées. Ceci est dû au fait que l'index de Tanimoto ne prend pas en compte l'absence d'entités dans la comparaison de deux molécules, et les petites molécules seront donc désavantagées par rapport aux grandes.

Quant à la considération ou non des absences de caractéristiques comme mesure de similarité, des discussions on été menées dans la communauté scientifique pour déterminer la validité et l'application chimique d'une telle affirmation.

Des propriétés analogues ont été remarquées pour d'autres mesures de similarité. Ainsi, d'autres différences importantes ont été relevées entre la distance Euclidienne et la distance de Hamming, d'une part, et les coefficients de Tanimoto, Dice et Cosinus, d'autre part: les premiers prennent en considération l'absence commune d'attributs comme une évidence de similarité contrairement aux derniers [Willett1998].

Un autre critère de classification et/ou d'évaluation des mesures de similarité est en relation avec le « principe de similarité des propriétés » [Johnson1990, Martin2002], principe qui a été déjà introduit dans le chapitre I.1. Selon ce critère, ce sont les indices ou coefficients qui expriment au mieux la relation entre la structure d'une molécule et ses propriétés, qui seront choisis. Des études menées sur ce sujet [Willett1986] ont montré que les coefficients de Tanimoto ou Cosinus ont de meilleures performances que les distances Euclidiennes ou de Hamming.

IV.2 Comparaisons intermoléculaires

On remarque que pour obtenir une mesure de similarité ou de diversité entre deux molécules réelles ou virtuelles, on utilise des **représentations** de ces molécules, représentations qui seront responsables dans une large mesure du succès ou de l'échec de la comparaison. Dans notre cas, des vecteurs moléculaires générés automatiquement sont tour à tour comparés en utilisant différents indices, coefficients ou distances.

Toutefois il est important de noter que toutes les mesures de similarité n'ont pas les mêmes propriétés vis-à-vis d'une même base de vecteurs représentatifs [Willett1998]. En fonction du nombre, de la nature et du type de données, on obtient alors des résultats qui peuvent être comparables, mais qui restent, en essence, différents ou complémentaires.

Dans les analyses de similarité et de diversité moléculaire qui utilisent comme descripteurs des informations chimiques groupées dans des entités définies basées sur des empreintes moléculaires, on observe que les molécules de grande taille, auront à priori beaucoup plus de chances que les molécules de petite taille d'avoir des entités en commun avec la molécule requête. C'est pour cela qu'il est recommandé d'introduire, dans le calcul de similarité ou de diversité, un facteur de normalisation en fonction de la taille des molécules. On empêche ainsi l'apparition d'écarts fictifs dans l'analyse de groupes de molécules hétérogènes. Cet effet est davantage prononcé quand on utilise des coefficients ou des distances qui prennent les absences d'information, comme une mesure de similarité (section IV.1).

Au sein de notre outil, les comparaisons moléculaires s'effectuent à travers l'examen des vecteurs représentatifs des molécules concernées. Cet ensemble de vecteurs moléculaires, pour un ensemble déterminé des molécules issues de la Query-TestDB, constitue l'outil de base pour effectuer des comparaisons intermoléculaires en utilisant des indices et des distances. Les informations contenues dans ce fichier sont le point de départ pour la génération d'un index de vecteurs, englobant les éléments à comparer, leurs propriétés, leurs caractéristiques, etc. Le fichier « index » de descripteurs a été construit et structuré en utilisant les mêmes principes que l'index de fragments de la FragDB. L'information dans les deux cas est abondante et comporte plusieurs niveaux de complexité. Dans le chapitre III.2.1, nous avons expliqué la procédure pour obtenir des vecteurs descripteurs de type structural pour chaque molécule analysée. Dans le chapitre II, nous avons montré également la structuration de cette information en utilisant des langages de marquage.

IV.2.1 Analyses de Similarité

La figure suivante montre, de manière schématique, le calcul de la similarité pour une paire de molécules représentées par ses vecteurs descripteurs 1 et 2. Ces deux vecteurs descripteurs peuvent être soit calculés directement par le logiciel avant d'effectuer la comparaison, soit chargés par l'outil dans leur format XML d'origine qui permet de récupérer toutes les informations nécessaires à l'opération.

Une fois les vecteurs descripteurs prêts, on procède à la recherche des fragments communs aux deux vecteurs. Ceci peut se faire de manière stricte, en comparant les noms codés des fragments concernés (homomorphisme structural), ou de manière plus floue en comparant seulement l'appartenance d'un fragment à une classe particulière car certaines informations du nom du fragment identifient l'appartenance de ses fragments à une même famille. On peut aussi ajouter des poids, pour que tous les fragments ne contribuent pas de la même manière au calcul de la similarité ou pour qu'une propriété particulière joue le rôle de discriminant principal.

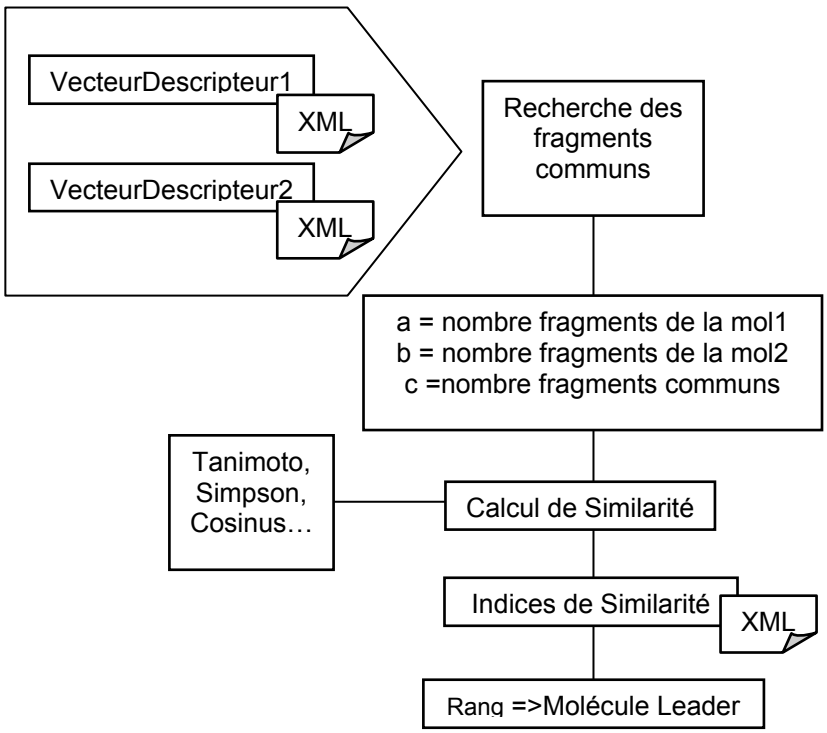


Figure IV.1 Schéma du calcul de la similarité entre deux molécules représentées par leurs vecteurs correspondants. L'étiquette « XML » identifie les données qui peuvent être traitées en entrée-sortie par l'outil dans un format XML.

Toutes ces considérations faites, on peut procéder au calcul de quantités *a, b* et *c*, nécessaires pour appliquer les formules de coefficients et/ou des indices déjà introduits dans la section IV.1 :

a : représente le **nombre des fragments** de la première molécule.
b : représente le **nombre des fragments** de la deuxième molécule.
c : est le **nombre des fragments** communs aux deux molécules.

L'indice de similarité calculé peut être enregistré dans un fichier XML accompagné des informations relatives aux molécules qui on servi de requête et de test. Ces informations nous permettront donc d'établir un classement (*rank*) ou liste ordonnée de molécules par rapport à leur similitude ou différence avec une ou plusieurs molécules requêtes. Dans la figure IV.1 on peut remarquer que nous utilisons principalement les indices de Cosinus, Tanimoto et Simpson pour effectuer des analyses de Similarité. La raison principale de ce choix réside dans le comportement

différent de ces trois mesures, ce qui nous permet d'avoir des regards différents sur le même groupe moléculaire. Par ailleurs, l'implémentation des autres coefficients ou distances signalés dans la table IV.1 ou définis par l'usager lui-même, ne présente pas de difficultés particulières.

IV.2.2 Calcul de la précision et du rappel *« recall »*

Les vecteurs représentatifs fournis par l'outil peuvent s'avérer incomplets, car il est certain que tous les fragments existant dans l'univers chimique n'ont pas été inclus dans la base de fragments. Notamment dans les cas des cycles, les combinaisons et permutations des différents substituants pour toutes les positions dans un cycle rendent la tâche quasi impossible.

Dans l'étape d'élaboration du programme, nous avons mis au point des fonctions visant l'évaluation de l'outil. Nous avons alors implémenté un module qui permet d'évaluer l'erreur et l'exactitude d'un vecteur représentatif déterminé, calculé par notre outil (dit vecteur « expérimental »), par rapport à un vecteur représentatif dit « théorique » car élaboré manuellement en observant minutieusement la structure de la molécule.

La **précision** d'une mesure est définie comme la quantité des correspondances dans toutes les réponses qui peuvent être possibles. Dans notre cas, ce sera le nombre de fragments identifiés par l'outil par rapport à ceux identifiés manuellement par un chimiste. Une mesure qui a 100% de précision indique que les résultats coïncident toujours avec les correspondances attendues. Mais rien n'est dit des non-correspondances. Ainsi, une molécule avec un vecteur descripteur qui contient 20 fragments, mais desquels seulement 8 correspondent aux 9 fragments attendus (les 12 autres étant des doublons, des fragments recouvrants, etc.), a une très grande précision, mais contient des fragments qui n'interviennent pas pour cette comparaison. Un des avantages (et limitations) de notre outil, est d'avoir une base prédéfinie de fragments. Ceci limite certainement les fragments détectés à ceux contenus dans la base, mais d'autre part, il n'y a pas de possibilité de détection des faux fragments (fragments mal définis, fragments redondants) car tous les fragments de la base ont été soigneusement sélectionnés et définis au préalable.

Le **rappel** ou *recall* vient compléter les résultats des calculs de précision. Le rappel est défini comme le nombre absolu de correspondances en prenant compte les non-correspondances. Ainsi une molécule avec un vecteur descripteur expérimental de 10 fragments aura 100% de rappel, si et seulement si, il y a 10 fragments attendus dans le vecteur descripteur théorique. Si le vecteur expérimental comporte plus ou moins de fragments, le rappel sera toujours inférieur car soit tous les fragments ont été identifiés mais il y a des fragments non attendus, soit tous les fragments n'ont pas été identifiés.

La ***F-measure*** [Van Rijsbergen1979], (rapport entre la précision et le rappel) vient compléter l'analyse des résultats. Pour analyser les résultats, avec une approche classe par classe, nous étudions la *F-mesure* de van Rijsbergen (1979) associée à chaque classe a priori : il s'agit de retrouver au mieux une classe experte dans l'ensemble de classes produites par un algorithme. Pour une analyse globale, nous pouvons également utiliser l'indice de Rand corrigé [Hubert1985] qui permet de comparer deux partitions. Pour les deux indices, une valeur de 0 correspond à une absence totale de correspondance entre la structure a priori et la structure obtenue, alors qu'une valeur de 1 indique une correspondance parfaite. Cette quantité permet donc de regrouper en un seul nombre les performances de l'outil (pour une classe donnée) pour ce qui concerne le rappel et la précision.

Les définitions des trois mesures dépendent de la structure vectorielle théorique construite manuellement pour évaluer l'outil « St » et de celle générée par l'outil, donc expérimentale « Sg ».

$$\mathrm{Pr}\,ecision = \frac{St \cap Sg}{Sg} \qquad Rappel = \frac{St \cap Sg}{St} \qquad \text{F - measure} = \frac{2 \times \left(\mathrm{Pr}\,ecision \times Rappel\right)}{\mathrm{Pr}\,ecision + Rappel}$$

La figure IV.2 illustre de manière schématique, le calcul de la précision, du rappel et de la *F-mesure* d'une molécule représentée par son vecteur descripteur.

D'autres mesures d'erreur sont aussi courantes:

L'**erreur absolue moyenne** (*mean absolute value*): pour chaque exemple, on calcule la différence entre la valeur théorique, et sa valeur expérimentale On divise ensuite la somme de ces erreurs par le nombre d'instances dans l'ensemble d'exemples.

Plus formellement :

– Soient $p_1, p_2, ..., p_n$ les valeurs correctement trouvées avec l'outil

– Soient $a_1, a_2, ..., a_n$ les valeurs attendues (théoriques)

Alors, l'Erreur absolue moyenne = $(|p_1 - a_1| + |p_2 - a_2| + \cdots + |p_n - a_n|)/n$

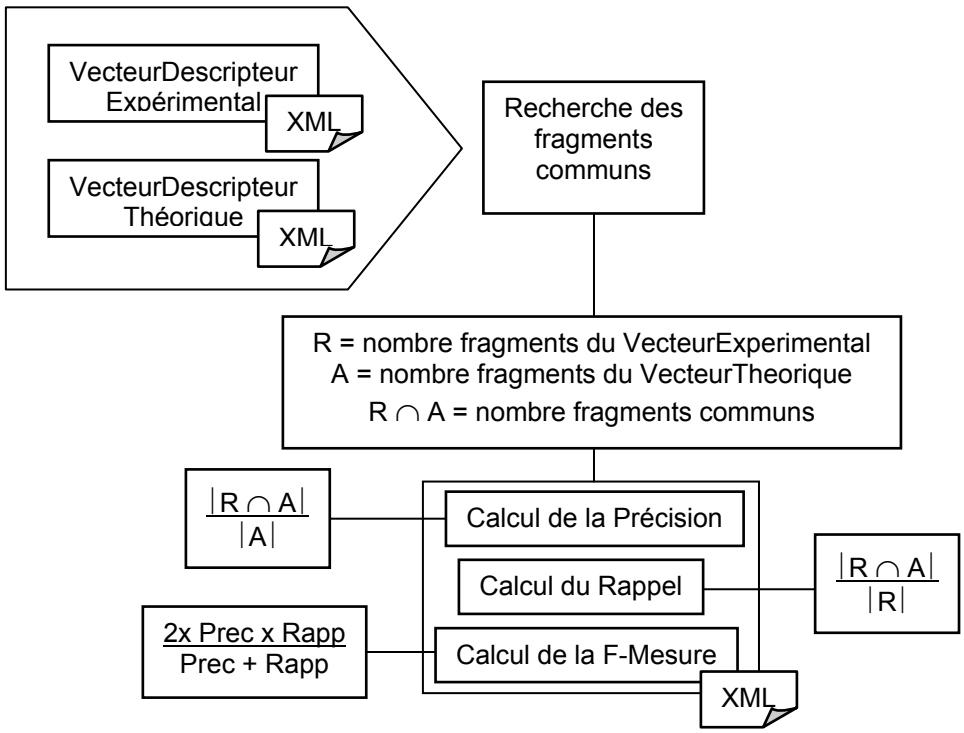


Figure IV.2 Schéma pour le calcul de la précision, du rappel et de la *F-Mesure* pour une molécule, en comparant le vecteur descripteur expérimental fourni par le logiciel et un vecteur descripteur théorique construit manuellement à partir de la structure moléculaire. Ceci a comme but la vérification de la fiabilité de résultats. L'étiquette « XML » identifie les données qui peuvent être traitées en entrée-sortie par l'outil dans un format XML.

La **Racine carrée de l'erreur quadratique moyenne** (*root mean-squared error*)**: c**ette mesure d'erreur concerne principalement la comparaison entre données expérimentales et valeurs réelles.

Avec les mêmes notations que ci-dessus, elle vaut:

Racine carrée de l'erreur quadratique moyenne = $\{[(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2] /n\}^{1/2}$

L'erreur quadratique avantage les vecteurs où il y a beaucoup de petits écarts, par rapport à ceux qui sont exacts presque partout, mais qui font de grosses erreurs en un petit nombre de points. Le fait de prendre la racine carrée permet de manipuler des quantités qui ont la même dimension que les valeurs à prévoir.

Toutes ces mesures d'erreur sont d'implémentation facile dans l'outil. Ceci dit, ce qui est moins facile est de disposer des vecteurs théoriques pour un grand groupe de molécules, en raison de leur construction essentiellement manuelle. Quant aux performances de l'outil, les résultats de l'évaluation du rappel, de la précision et de la F-mesure sont présentés dans le chapitre 5.

IV.3 Les différents niveaux de comparaison

Pour effectuer les différents niveaux de comparaison, on procède à l'analyse des vecteurs descripteurs et des fragments les composant.

D'abord, la comparaison stricte ou exacte des noms de fichier des fragments trouvés nous permet d'effectuer des homomorphismes entre les fragments de la molécule détectés par l'outil et ceux de la FragDB.

Si cette recherche stricte échoue, on passe à l'analyse des informations codées par le nom de fichier des fragments trouvés. On extrait la connaissance, d'ordre structural, de classes de molécules et des propriétés choisies, incorporée dans les vecteurs descripteurs. Cette information nous permet d'effectuer des comparaisons non exactes sur des critères particuliers. Par exemple, l'appartenance ou non d'un fragment détecté à une famille des fragments (les critères de définitions des « familles » sont d'ordre structural principalement: ainsi une amine tertiaire pourrait être comparé à une secondaire ou primaire). Les niveaux de « flou » peuvent varier d'un fragment à un autre: ainsi, pour certains fragments, seul la comparaison exacte sera possible (dû surtout à sa structure), alors que pour d'autres fragments, des « flous structuraux » pourront englober des familles de 4 ou plus

des fragments « équivalents ». En suivant cette méthode, on a plus de chances de décrire la molécule en totalité.

Parfois le logiciel ne sera pas capable de générer le fragment correspondant et d'effectuer la comparaison, soit parce que le fragment n'est pas dans la base, soit parce que les informations moléculaires sont insuffisantes pour effectuer l'algorithme d'Ullmann. Et parfois, le logiciel proposera par induction des fragments flous que l'on n'avait pas prévu au début.

On peut aussi ajouter des poids, au moment de la comparaison fragment-fragment pour paramétrer l'importance relative des structures et des propriétés.

L'importance d'avoir différents types et niveaux de comparaison réside dans la possibilité d'effectuer de multiples analyses en fonction de la complexité et de la nature des données moléculaires.

Quatre possibilités pour l'analyse de la Similarité et de la Diversité moléculaires sont offertes dans notre approche. Celles-ci sont représentées dans la figure IV.3

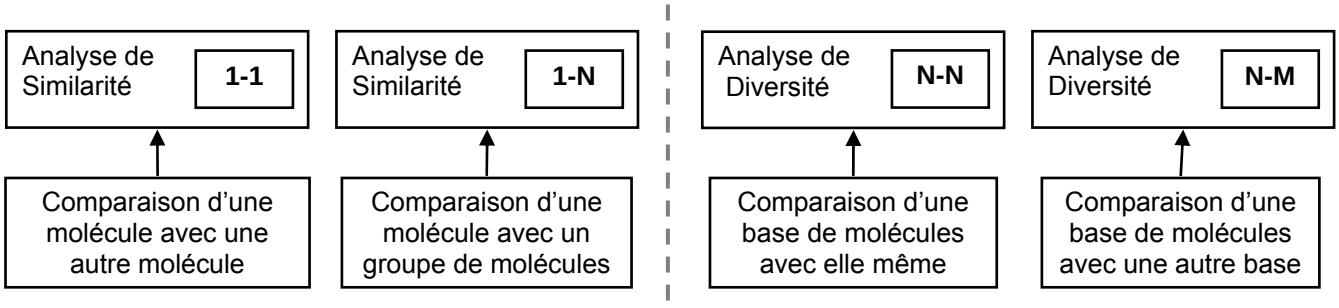

Analyse de Similarité
1-1
Analyse de Similarité
1-N
Analyse de Diversité
N-N
Analyse de Diversité
N-M
Comparaison d'une molécule avec une autre molécule
Comparaison d'une molécule avec un groupe de molécules
Comparaison d'une base de molécules avec elle même
Comparaison d'une base de molécules avec une autre base

Figure IV.3. Analyses de similarité et de diversité proposées pour notre outil.

Le premier cas repose sur un calcul unique de la similarité entre la molécule 1 et la molécule 2 (cas 1-1). On peut aussi calculer la similarité d'une molécule avec une base de N molécules (cas 1-N). Le troisième cas consiste en un calcul de la diversité interne d'une base de molécules données (cas N-N) au travers des techniques expliquées auparavant. La quatrième possibilité consiste à effectuer

un calcul de la diversité d'une base de molécules 1 par rapport à une base de molécules 2 (cas N-M). Les deux dernières analyses produiront des **matrices** de Similarité/Diversité, de taille N × N et N × M respectivement.

D'une part, l'analyse de la similarité moléculaire fournit une méthode simple et courante pour le criblage virtuel et elle est à la base des méthodes de *clustering*. D'autre part, l'analyse de la diversité moléculaire explore la façon dont les molécules peuplent un espace structural déterminé, et elle est à la base de beaucoup d'approches pour la conception des bibliothèques combinatoires et le choix de leurs composés. Le choix d'un espace métrique optimal qui représente correctement la diversité structurale, ainsi que des descripteurs qui expriment la réalité chimique, sont déterminants dans l'efficacité du modèle.

Mais, comment construire des sous-ensembles diversifiés de bases de données chimiques, par exemple, pour l'inclusion dans un programme de criblage biologique ou la construction des bibliothèques combinatoires ? Le but est donc d'identifier des sous-ensembles avec un maximum de différences selon des critères structuraux 2D, de forme 3D ou d'activité. Puisque l'identification du sous-ensemble le plus divers exige l'utilisation d'un algorithme combinatoire qui considère tous les sous-ensembles possibles à partir d'un ensemble de données déterminé, l'identification du sous-ensemble de diversité moléculaire maximale prendra beaucoup de temps ! D'où l'intérêt d'automatiser les processus qui mènent à identifier un groupe diversifié des molécules dans des grandes bases de données chimiques.

Pour effecteur cette tache, un algorithme assez courant consiste à prendre une molécule au hasard et à la placer dans un sous-groupe « divers ». On cherche ensuite, dans la base d'origine, la molécule la plus dissimilaire à cette molécule et on continue ainsi de suite jusqu'à n'avoir plus de molécules dissimilaires dans la base originale [Willett1987]. Différents critères de dissimilarité peuvent être utilisés ainsi que différents seuils de mesures de dissimilarité (1 - coefficient de Similarité), pour calculer les différences entre les molécules. On pourra obtenir ainsi des sous-groupes différents à partir de la même molécule « graine » choisie pour commencer l'algorithme [Willett1996].

Deux niveaux de comparaison sont également proposés à ce jour:

- Le premier niveau consiste à prendre en compte seulement l'information structurale des molécules.

- Le deuxième niveau prend en compte les propriétés physicochimiques en plus des informations structurales des molécules.

Ces niveaux sont représentés dans la figure IV.4. Dans la section IV.3 nous allons détailler avec un exemple les niveaux de comparaison.

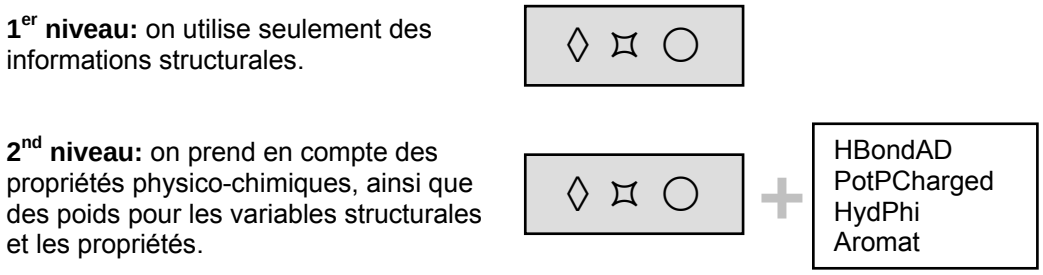

1er niveau: on utilise seulement des informations structurales.
2nd niveau: on prend en compte des propriétés physico-chimiques, ainsi que des poids pour les variables structurales et les propriétés.
HBondAD
PotPCharged
HydPhi
Aromat

Figure IV.4. Niveaux de comparaison proposés pour effectuer des analyses de similarité et de diversité. Les symboles dans le cartouche, représentent des fragments moléculaires.

L'importance d'offrir autant de niveaux de comparaison réside dans l'étendue des applications possibles de l'outil. Les combinaisons des calculs augmentent la diversité d'usage de l'outil. De requêtes bibliographiques (nettement structurales), à la recherche de pharmacophores ou des molécules ayant des propriétés particulières, les applications sont donc multiples.

IV.3.1 Comparaison exclusivement structurale

Nous partons de deux vecteurs résultants V et V', qui correspondent à deux molécules différentes. On considère que chaque vecteur contient respectivement n et n' fragments, avec la condition que $n \geq n'$. La lettre « f » dénote les fragments qui décrivent ces deux molécules V et V' dans la formule :

$$V = (f_1, f_2, ..., f_n)$$
$$V' = (f'_1, f_2', ..., f_n') \quad (1)$$

En général, comparer deux vecteurs se résume à retrouver la distance qui les sépare dans un espace défini. Beaucoup de mesures de distance, de coefficients et d'indices existent aujourd'hui. Dans notre outil, nous présentons une sélection des mesures de comparaison, avec comme but de pouvoir effectuer un maximum d'analyses en fonction des données moléculaires et des problèmes posés. Une fois la mesure de comparaison choisie, on peut structurer les résultats dans des fichiers XML pour faciliter l'accès à l'information. Nous allons considérer en conséquence que pour comparer les vecteurs *V* et *V'* on utilise une distance « *D(V,V')* », qui remplit les conditions suivantes (pour plus de détails, voir le tableau IV.1) :

1) $D(V,V') = 0$ quand les vecteurs (les molécules) sont totalement différents.

2) $D(V,V') = 1$ quand les vecteurs (les molécules) sont identiques.

3) $D(V,V') = D(V',V)$ car la mesure de comparaison est symétrique.

La mesure de comparaison prendra en compte l'approche par sous-structure déjà expliquée. L'importance de chaque fragment (ou de ses familles structurales) peut être paramétrée en utilisant des « poids structuraux » choisis par l'usager. Si aucun poids n'est précisé, tous les fragments auront la même importance vis-à-vis de la formule de similarité ou diversité. Ainsi, les poids ont une valeur par défaut de « 1 », et peuvent être paramétrés avec des valeurs allant de zéro (ne pas prendre en compte cette structure) à deux (structure très importante pour le calcul). Une fois pris en compte les « poids structuraux », les vecteurs *V* et *V'* auront l'apparence suivante (2):

$$V = (f_1 \times w_1, f_2 \times w_2, ..., f_n \times w_n)$$
$$V' = (f_1' \times w_1', f_2' \times w_2', ..., f_n' \times w_n') \quad (2)$$

Où « f_i » est un fragment descripteur de molécules et « w_i » son poids structural. Les vecteurs de l'équation (2) fournissent l'information de base qui sera utilisée par notre outil

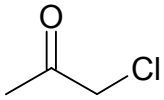
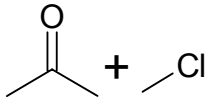
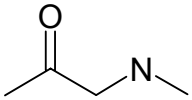
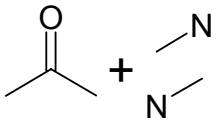
Nom de la Molécule	Structure molécule	Découpage molécule	Fragments du Vecteur Représentatif
Molécule *V :* 1-Chloro-propan-2-one			< f_1: AGCC-014R, f_2: ANSX-000X; 2 >
Molécule *V' :* 1-Methylamino-propan-2-one			< f_1: AGCC-014R, f_3: ANSZ-000Z, f_4: ANSZ-000Z ; 3 >>

Figure IV.5 Construction du vecteur descripteur pour deux molécules données. Présentation des molécules, découpage, puis construction du vecteur représentatif.

Prenons par exemple la comparaison entre les molécules *V* et *V'* du tableau IV.5 effectué sans modification des poids et utilisant comme mesure de similarité l'index de Tanimoto. Cette mesure pour des valeurs continues obéit à la formule suivante :

$$S_T = \frac{c}{a+b-c} \tag{3}$$

Où :

$$\begin{aligned} a &= \sum_i w_i f_i^{\,2} \\ b &= \sum_i w_i' f_i'^{\,2} \\ c &= \sum_i w_i f_i f_i' \end{aligned} \tag{4}$$

a, b et *c*, représentent respectivement la somme des fragments de la première molécule (f_1 et f_2), de la deuxième molécule (f_1, f_3 et f_4) et le nombre des fragments communs (f_1). On peut alors calculer

sans aucune difficulté la similarité entre les molécules *V* et *V'* en utilisant l'équation (3) et les valeurs de la figure IV.5 :

$$S_T = \frac{c}{a+b-c} = \frac{1}{2+3-1} = 0{,}25 \qquad (5)$$

Si l'usager décide plus tard de modifier les poids des fragments comme l'indique le tableau de la figure IV.6, la mesure de similarité entre les molécules *V* et *V'* sera modifié.

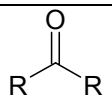
Fragment	Structure	Poids Cas 1	Poids Cas 2
f_1		1	2
f_2	R—X	2	0
f_3	R—z	1	1
f_4	R—z	1	1

Figure IV.6 Deux modifications possibles de poids des molécules. Dans le premier cas le poids correspondant au carbonyle a été diminué de moitié, dans le second cas le même fragment subit une augmentation de son poids général, par rapport aux autres fragments.

Les valeurs « *a, b, c* » pourront être à nouveaux calculés avec les équations (4). Tous les fragments ne contribueront pas de la même manière, et seront modifiés en fonction de leur poids structural.
Pour le cas 1, $a' = 3$; $b' = 3$ et $c' = 1$. Tant que pour le cas 2, $a'' = 2$; $b'' = 4$ et $c'' = 2$. On peut alors recalculer la similarité entre les molécules *V* et *V'* en utilisant l'équation (3) et les valeurs de la figure IV.6. On obtient ainsi que la mesure de similarité en utilisant des poids structuraux (équation 6) est modifiée par rapport au calcul de base (équation 5).

$$S_T' = \frac{c'}{a'+b'-c'} = \frac{1}{3+3-1} = 0{,}20$$
$$S_T'' = \frac{c''}{a''+b''-c''} = \frac{2}{2+4-2} = 0{,}50 \qquad (6)$$

Les résultats sont logiques puisque quand un fragment commun a un poids structural plus important que les autres, la similarité entre les deux molécules augmente ($S_T'' = 0{,}50$ par rapport à $S_T' = 0{,}25$). Et inversement, si son poids est moins important ($S_T' = 0{,}20$ par rapport à $S_T = 0{,}25$).

Il ne faut pas oublier qu'un vecteur n'est après tout qu'un modèle très simplifié d'une molécule, et que l'on peut avoir des résultats inattendus au moment de les comparer. Par exemple, si on part de l'idée qu'une molécule est égale à un vecteur, on peut assurer que les mesures de comparaison seront uniques entre deux molécules différentes. Mais, selon le modèle employé ici pour générer des vecteurs représentatifs, le fait de ne pas avoir considéré la disposition spatiale (3D) de molécules ou les isomères optiques et chiraux, a comme conséquence probable l'apparition d'un même vecteur représentatif pour plusieurs molécules. La mesure de similarité entre les deux vecteurs moléculaires sera donc moins représentative de la réalité chimique.

IV.3.2 Comparaison reposant sur la structure et les propriétés des molécules

Le deuxième niveau de calcul proposé inclut les propriétés physicochimiques (p_i). Comme nous l'avons indiqué auparavant, l'usager peut paramétrer l'importance des propriétés choisies en utilisant des « poids propriété » (v_i).

Dans la section IV.3.2 nous avons montré comment paramétrer certains types de sous-structures (*fi*) en utilisant des « poids structuraux » (w_i). La valeur par défaut de tous les poids est égale à l'unité et permet de donner la même importance à toutes les propriétés et à toutes les structures proposées à l'usager. A l'instar de (2) on peut donc définir de manière générique l'effet des poids de propriétés sur les molécules *V* et *V'* de la manière suivante.

$$V = \left(\left(\sum_j p_{1j} v_j \right) \times w_1, \ldots, \left(\sum_j p_{nj} v_j \right) \times w_n \right)$$
$$V' = \left(\left(\sum_j p'_{1j} v_j' \right) \times w_1', \ldots, \left(\sum_j p'_{n'j} v_j' \right) \times w_{n'}' \right) \quad (7)$$

Où pij est la $j^{ème}$ propriété du $i^{ème}$ fragment des deux molécules V et V', v_j le poids propriété et w_i le poids structure correspondants. Pour chaque fragment i, avec j propriétés : $f_i = \left(\sum_j p_{ij} v_j \right) \times w_i$ on peut regrouper l'information concernant les propriétés physicochimiques, les structures et ses poids, dans un élément « e_i » avec la structure suivante.

$$e_i = \sum_j p_{ij} v_j \frac{w_i}{\sum_j p_j} \quad (8)$$

Suivant les nouvelles modifications, pour les valeurs « *a, b, c* » de l'équation (4), les fragments ne contribueront pas de la même manière, et pourront être redéfinis en forme d'entités :

$$a = \sum_{i=1}^{n} e_i, \quad b = \sum_{j=1}^{n'} e_j, \quad c = \sum_{k=1}^{\min(n,n')} e_k \quad (9)$$

Où e_k représente les éléments en commun entre e_i et e_j

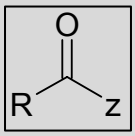
Un échantillon de propriétés a déjà été donné dans la figure II.17 et dans le code II.7 au moment de la description des fragments structurés en utilisant les langages de marquage. Dans ces mêmes

figures on montre comment, à partir des données structurales contenues dans le fichier MOL, on peut déduire automatiquement des valeurs assignées pour certaines propriétés physicochimiques.

Les propriétés choisies fournissent de l'information par rapport à la polarisation de la molécule, à son caractère aromatique, à la capacité de donner ou d'accepter des atomes H, etc. Les poids associés auront par défaut une valeur de « 1 », et peuvent être paramétrés avec des valeurs allant de zéro (ne pas prendre en compte cette propriété) à deux (propriété très importante pour le calcul). Tous les poids (structuraux ou propriétés) sont normalisés avant d'effectuer le calcul. Ces poids (v_i et w_i) choisis par l'usager jouent le rôle de valeurs de pondération ou de coefficients de normalisation. Ainsi, quand les deux vecteurs à comparer n'ont pas la même taille, la normalisation des poids (tant pour le plus grand comme pour le plus petit d'entre eux) a comme but de ne pas fausser les résultats de comparaison.

Reprenant la structure de données simplifiées pour FragDB montrée dans le code II.7, et en faisant quelques modifications pour rendre plus claires les données qui nous intéressent (clefs de recherche et propriétés physicochimiques), on obtient le code IV.1.

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<!-- Sample of Index.xml : Data Structure for FragDB  -->
<Index>
      <File name="AGCZ-014Z.mol">
            <PositionList>
                  ....
            </PositionList>
            <Keys>
                  <Key name="FID" value="014Z"/>
                  <Key name="FAtomSum" value="4"/>
                  ....
            </Keys>
            <Properties>
                  <Property name = "HBondAD" value = "1"/>
                  <Property name = "Aromat" value = "0"/>
                  <Property name = "Polar" value = "1"/>
                  ....
            </Properties>
      </File>
....
</Index>
```



Code IV.1 Index-XML : Structure de données simplifiée pour FragDB

On observe que les données correspondant aux propriétés physicochimiques et aux clefs de recherche ne sont pas seulement facilement repérables, mais elles sont aussi parfaitement structurées. Ainsi, les vecteurs de l'équation (7) reprendront cette information qui sera ensuite paramétrée avec les poids choisis par l'usager. Ces données sont utilisées par notre outil pour effectuer des mesures de comparaison en utilisant une formule *D(1,2)* qui pourra être choisie parmi une liste de mesures disponibles.

Prenons à nouveau, par exemple, les deux molécules montrées dans la figure IV.5. On considère que l'algorithme d'Ullmann a détecté les fragments (f_1, f_2, f_3 et f_4). Si chaque fragment (f_i) a 3 propriétés (p_{ij}) associées, on obtient à l'instar de (7), le groupe d'équations (10).

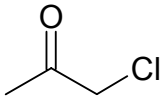
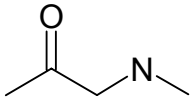
Nom de la Molécule	Structure molécule	Fragments du Vecteur Représentatif
Molécule *V :* 1-Chloro-propan-2-one		<f_1: AGCC-014R, f_2: ANSX-000X; 2 >
Molécule *V' :* 1-Methylamino-propan-2-one		< f_1: AGCC-014R, f_3: ANSZ-000Z, f_4: ANSZ-000Z ; 3 >>

Figure IV.7 : Vecteur résultat pour les deux molécules de la figure IV.5.

$$V = \left(\left(p_{11}v_1 + p_{21}v_2 + p_{31}v_3\right)w_1, \left(p_{12}v_1 + p_{22}v_2 + p_{32}v_3\right)w_2\right)$$
$$V' = \left(\left(p_{11}v_1 + p_{21}v_2 + p_{31}v_3\right)w_1, \left(p_{13}v_1 + p_{23}v_2 + p_{33}v_3\right)w_3, \left(p_{14}v_1 + p_{24}v_2 + p_{34}v_3\right)w_4\right) \quad (10)$$

On remarque que les deux vecteurs ont un seul fragment en commun. Une analyse de premier niveau comme celle présentée dans la section IV.3.1 et qui compare exclusivement les structures des molécules, donnera comme mesure de similarité simple: 0.25 (rappelons que nous avons obtenu 0,20 pour le cas 1 et 0,50 pour le cas 2).

Si l'on effectue une analyse de deuxième niveau, sur le même groupe de molécules et si l'on prend en compte les propriétés des fragments constitutifs, la valeur de similarité simple pourra varier en fonction des poids des propriétés des fragments.

Par exemple, on peut considérer les valeurs suivantes pour les propriétés des molécules montrées dans la figure IV.7 :

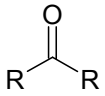
Fragment	Structure	Propriété 1 : Aromaticité	Propriété 2 : Polarisation	Propriété 3 : Accepteur H
f_1		1	1,5	2
f_2	R—X	1	2	1
f_3	R—z	1	1	1
f_4	R—z	1	1	1
Poids des propriétés		0	2	1

Figure IV.8 Table des valeurs des propriétés pour les fragments des molécules *V* et *V'*. Des valeurs possibles de poids pour les propriétés sont indiquées à la fin. La valeur par défaut dans tous les cas est égale à 1.

L'usager choisit ensuite les valeurs / poids à assigner aux fragments et aux propriétés physicochimiques, en fonction de l'importance ou de la pertinence qu'ils présentent pour son problème ou pour le calcul de la similarité / diversité moléculaire.

Pour les fragments, reprenons les poids structuraux du cas 2 déjà présenté dans la figure IV.6. Pour les propriétés, imaginons que les poids des propriétés choisis par l'usager correspondent à ceux de la figure IV.8. Dans cet exemple, la présence d'un carbonyle et une haute polarisabilité sont les

critères qui ressortent du choix des poids, pour le calcul de la mesure de similarité entre les deux molécules.

La contribution de chaque fragment sera regroupée sous forme d'entité e_i (formule 8) qui permettront de calculer les valeurs *a, b* et *c,* pour le calcul de la mesure de similarité. En utilisant les valeurs de la figure IV.6, IV.8 et les équations (8), (9) et (10), on obtient pour notre exemple, le résultat suivant:

$$
\begin{aligned}
e_i &= \frac{\left((p_{1i} \times v_1) + (p_{2i} \times v_2) + (p_{3i} \times v_3)\right)}{\sum_{k=1}^{3} p_k} \times w_i \\
e_1 &= \frac{\left((1 \times 0) + (1{,}5 \times 2) + (2 \times 1)\right)}{4{,}5} \times 2 = 2{,}22 \\
e_2 &= \frac{\left((1 \times 0) + (2 \times 2) + (1 \times 1)\right)}{4} \times 0 = 0 \\
e_3 &= \frac{\left((1 \times 0) + (1 \times 2) + (1 \times 1)\right)}{3} \times 1 = 1 \\
e_4 &= \frac{\left((1 \times 0) + (1 \times 2) + (1 \times 1)\right)}{3} \times 1 = 1
\end{aligned}
\qquad (11)
$$

Suivant les nouvelles modifications, pour les valeurs « *a, b, c* » de l'équation (3) on obtient:

$$
\begin{aligned}
a &= 2{,}22 + 0 = 2{,}22 \\
b &= 2{,}22 + 1 + 1 = 4{,}22 \\
c &= 2{,}22
\end{aligned}
\qquad (12)
$$

On peut alors calculer la similarité entre les molécules *V* et *V'* en utilisant l'équation (3) et les valeurs de (12):

$$
S_T = \frac{c}{a + b - c} = \frac{2{,}22}{2{,}22 + 4{,}22 - 2{,}22} = 0{,}52 \qquad (13)
$$

Finalement, la mesure de similarité calculée en utilisant des poids de structures et de propriétés est de **0.52**, au lieu de 0.25 au premier niveau d'analyse, et de 0.50 quand sont utilisés seulement les poids structuraux. Si on prend donc en compte les propriétés physicochimiques des molécules *V* et *V'*, celles ci sont plus similaires, que si l'on prend en compte seulement leur structure, car un des fragments commun aux deux structures, à une importance plus grande à l'égard de ses propriétés attachées.

Conclusion

Dans ce chapitre, nous avons étudié les indices de similarité et de diversité dans un cadre de criblage de molécules. De très nombreuses représentations existent aujourd'hui pour exprimer la comparaison entre deux descripteurs moléculaires, ou pour établir une distance entre deux objets dans un espace déterminé. Dans ce chapitre, nous avons expliqué l'usage des ces mesures et l'automatisation de ces processus comme des conditions vitales pour le traitement des grandes bases de données. Quatre possibilités pour l'analyse de la Similarité et de la Diversité ont été présentées dans notre approche. Plusieurs types de calculs sont ainsi proposés.

Références

[Dixon1999] Dixon, S.L., Koehler, R.T., *The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries*, J. Med. Chem., 42 (1999) 2887-2900.

[Johnson1990] Johnson, A.M., Maggiora, G.M. (Eds.) Concepts and Applications of Molecular Similarity, John Willey & Sons, New York, Inc. 1990.

[Holliday2002] Holliday, J.D., Hu, C.Y., Willett, P., *Grouping of coefficients for the calculation of Inter-molecular similarity and dissimilarity using 2D fragment Bit-Strings,* Comb. Chem. High Throughput Screening, 5 (2002) 155-166.

[Holliday2003] Holliday, J.D., Salim, N., Whittle, M., Willett, P., *Analysis and display of the size of chemical similarity coefficients,* J. Chem. Inf. Comput. Sci., 43 (2003) 819-828.

[Hubert1985] Hubert, L., Arabie, P. *Comparing partitions.* Journal of Classification 2, 193–218 (1985).

[Lajiness1997] Lajiness M.S., *Dissimilarity-based compound selection techniques*, Persp. Drug Discuss. Design, 7/8 (1997) 65-84.

[Martin2002] Martin, Y.C., Kofron, J.L., Traphagen, L.M. *Do structurally similar molecules have similar biological activity?,* J. Med. Chem., 45 (2002) 4350-4358.

[Pearlman1999] Pearlman, R.S., *Novel Software Tools for addressing Chemical Diversity,* Network Science (1999). Disponible sur: http://www.netsci.org/Science/Combichem/feature08.html

[SimWeb] Information disponible sur : http://pro.chemist.online.fr/cours/similarite.htm

[Whittle2004] Whittle, M., Gillet, V., Willett, P., *Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a Comparison of Similarity Coefficients*. J. Chem. Inf. Comput. Sci., 44 (2004) 1840-1848.

[Van Rijsbergen1979] Van Rijsbergen, C.J., Information Retrieval (second ed.). London. Butterworths, 1979.

[Willett1986] Willett, P., Winterman, V., Bawden, D., *Implementation of Nearest Neighbor Searching in an Online Chemical Structure Search System*, J. Chem. Inf. Comput. Sci., 26 (1986) 36-41.

[Willett1986] Willett, P., Winterman, V.A. *Comparison of some measures for the determination of intermolecular structural similarity measures,* Quant. Struct. -Act. Relat., 5 (1986) 18-25.

[Willet1987] Willett, P. Similarity. Clustering in Chemical Information Systems. Letchworth: Research Studies Press, 1987.

[Willett1996] Willett, P., *Molecular diversity techniques for chemical databases*. Information Research, 2 (1996). Information disponible sur: http://informationr.net/ir/2-3/paper19.html

[Willett1998] Willett, P., Barnard, J.M., Downs, G.M., *Chemical Similarity Searching*, J. Chem. Inf. Comput. Sci., 38 (1998) 983-996.

CHAPITRE V.

PRESENTATION ET

ANALYSE DES RESULTATS

V.1 Analyse de type 1-N
V.1.1 Résultats avec la base « Zinc »
V.1.2 Résultats avec la base « Random »
V.1.3 Comparaison des indices selon le rang
V.1.3.1 Graphiques de comparaison d'indices avec la base « Zinc »
V.1.3.2 Graphiques de comparaison d'indices avec la base « Random »
V.1.4 Comparaison des indices selon la complexité
V.1.4.1 Graphiques de comparaison d'indices avec la base « Zinc »
V.1.4.2 Graphiques de comparaison d'indices avec la base « Random »
V.2 Analyse de type N-N
V.2.1 Résultats avec la base « Zinc »
V.2.2 Résultats avec la base « Random »
V.2.3 Aperçu des résultats structurés et présentés avec XML
V.3 Evaluation de l'outil
V.3.1 Précision, rappel, et F-mesure, pour la base « Zinc »
V.3.2 Etude des faux isomorphismes pour des mesures de similarité N-N
V.3.3 Limites et avantages de l'outil

CHAPITRE V.

PRESENTATION ET ANALYSE DES RESULTATS

Tout au long du présent manuscrit, nous avons expliqué le fonctionnement de MolDiA : un système de criblage virtuel pour l'analyse de la similarité et de la diversité des bases de données moléculaires. Cet outil repose sur une nouvelle conception de diversité qui inclut des informations structurales et des propriétés physicochimiques. Dans ce chapitre nous effectuerons des analyses de similarité et de diversité en prenant en compte des critères divers, tels la taille des molécules ou la nature de la base. Les résultats seront groupés selon le type d'analyse effectué : 1-N ou N-N, en utilisant des différents niveaux de comparaison avec différentes bases moléculaires.

Les résultats présentés dans cette section ont été obtenus en utilisant deux bases moléculaires différentes:

- la base « Zinc » composée d'un échantillon de 34 molécules extraites de la base de molécules ZINC [ZincWeb]. ZINC est une base gratuite et en ligne qui contient des drogues et des molécules mises à disposition pour effectuer des études de criblage virtuel.
- la base « Random » composé d'un échantillon de 77 molécules choisies de manière aléatoire « *random* » et fournies par le Dr. Markus Meringer [Meringer2006]. Cette base est composée de molécules relativement petites (de 3 à 13 atomes de carbone) par rapport aux molécules de la base « Zinc ».

V.1 Analyse de type 1-N

Une analyse 1-N consiste à comparer une molécule cible avec une base de molécules test. Les résultats des mesures de similarité sont présentés sous forme de tableaux qui peuvent être classés ou manipulés en fonction des besoins.

V.1.1 Résultats avec la base « Zinc »

Lors de l'analyse 1-N avec les molécules de la base « Zinc », nous avons utilisé 4 molécules cibles différentes, chacune d'entre elles nous permettant de tester une propriété ou une caractéristique déterminée de l'outil.

- Pour tester la détection et le traitement des systèmes cycliques aromatiques et hétérocycliques, ainsi que des groupes fonctionnels, on a utilisé deux molécules : Query1Z2 et Query2Z2, relativement petites (10-11 atomes). Ces molécules nous permettront également de tester la capacité de l'outil à retrouver des groupes fonctionnels courants.
- Pour tester la capacité de l'outil à traiter les cycles liés, ainsi que les molécules contenant des hétéroatomes dans des chaînes acycliques, nous avons utilisé une molécule (Query1Z3) plus grande et complexe (20 atomes), avec un système bi-cyclique.
- Finalement, pour évaluer l'efficacité des algorithmes développés pour l'utilisation des atomes génériques (Q, Z, R, M,… pour plus de détails sur les atomes génériques, voir le chapitre II) dans les requêtes moléculaires, on a utilisé Query3Z.mol. Dans cette molécule, les atomes utilisés (-Q, -X) ne font pas partie des systèmes cycliques. Toutefois, des requêtes comprenant des atomes génériques dans les cycles moléculaires sont tout à fait réalisables.

Des classements « top10 » pour tous les indices, sont présentés dans les tableaux V.1-V.4. L'ordre des mesures correspond à celui de l'index de Simpson. Si on fait un graphique « indice de similarité *vs* noms de molécules », on peut construire pour chaque tableau de résultats (Tanimoto, Simpson et Cosinus) un graphe de la tendance de la base par rapport à chaque index. Quelques molécules composant la base « Zinc » sont montrées à la fin de ce chapitre.

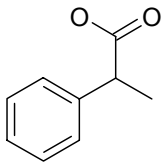


Tableau V.1 Résultats avec cible : Query1Z2.mol

Filename	Tanimoto	Cosinus	Simpson
../mol/ZincQueries/Query1Z2.mol	1.000000	1.000000	1.000000
../mol/ZincTest/zinc_10.mol	0.666667	0.800000	0.800000
../mol/ZincTest/zinc_189.mol	0.400000	0.596285	0.800000
../mol/ZincTest/zinc_38.mol	0.363636	0.565685	0.800000
../mol/ZincTest/zinc_57.mol	0.333333	0.539360	0.800000
../mol/ZincTest/zinc_58.mol	0.363636	0.565685	0.800000
../mol/ZincTest/zinc_1146.mol	0.333333	0.516398	0.666667
../mol/ZincTest/zinc_888.mol	0.333333	0.516398	0.666667
../mol/ZincTest/zinc_1037.mol	0.214286	0.387298	0.600000
../mol/ZincTest/zinc_1527.mol	0.214286	0.387298	0.600000
../mol/ZincTest/zinc_370.mol	0.300000	0.474342	0.600000

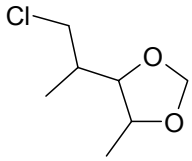
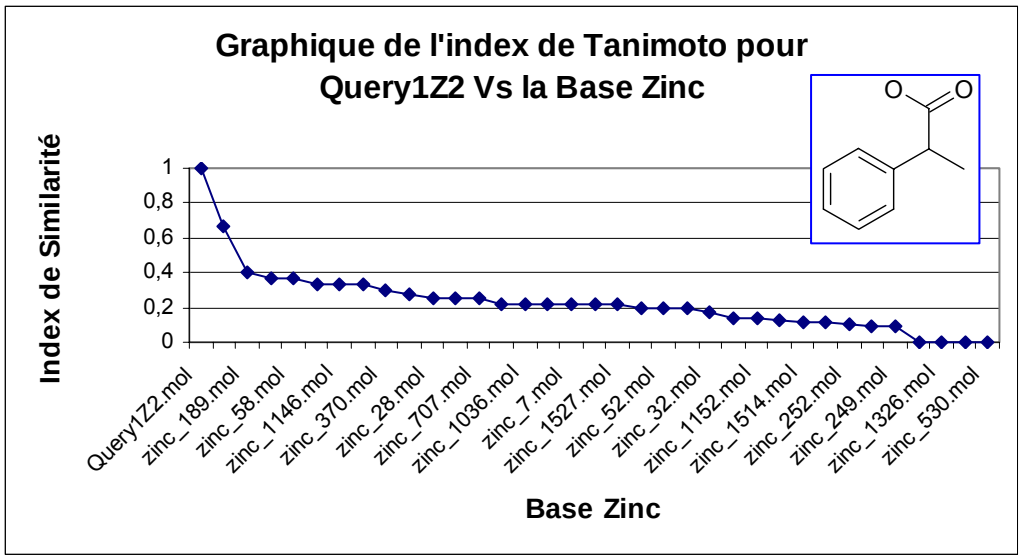
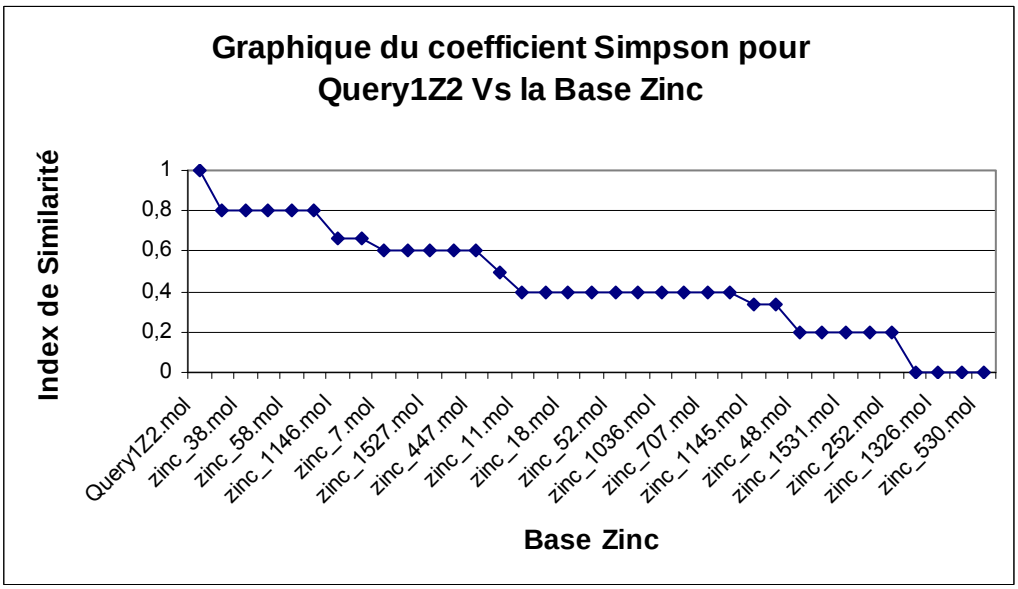


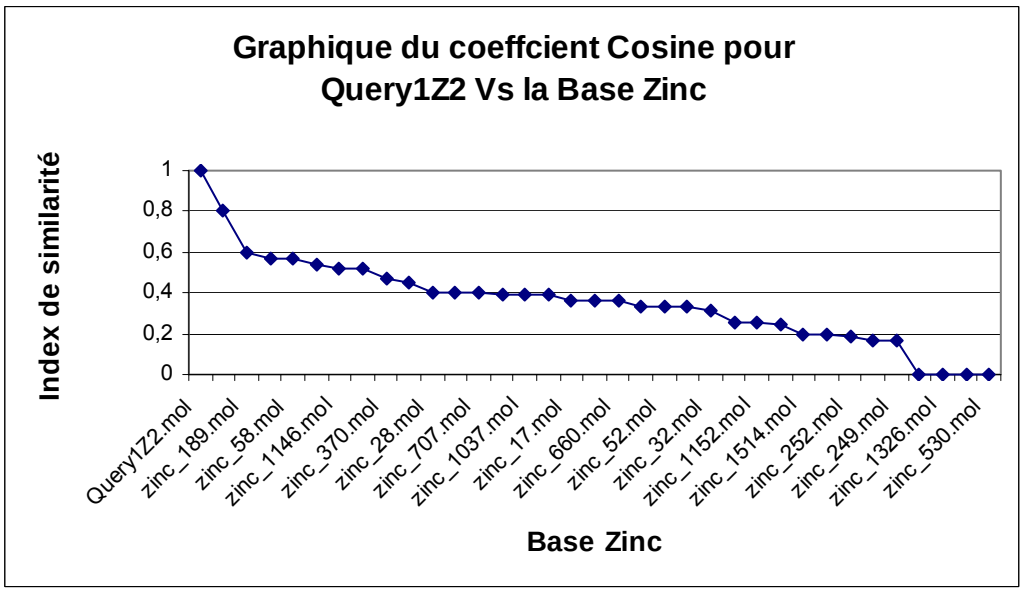
Tableau V.2 Résultats avec cible: Query2Z2.mol

Filename	Tanimoto	Cosinus	Simpson
../mol/ZincQueries/Query2Z2.mol	1.000000	1.000000	1.000000
../mol/ZincTest/zinc_18.mol	0.444444	0.617213	0.666667
../mol/ZincTest/zinc_38.mol	0.333333	0.516398	0.666667
../mol/ZincTest/zinc_1037.mol	0.285714	0.471405	0.666667
../mol/ZincTest/zinc_1527.mol	0.285714	0.471405	0.666667
../mol/ZincTest/zinc_189.mol	0.363636	0.544331	0.666667
../mol/ZincTest/zinc_447.mol	0.363636	0.544331	0.666667
../mol/ZincTest/zinc_888.mol	0.285714	0.471405	0.666667
../mol/ZincTest/zinc_28.mol	0.375000	0.547723	0.600000
../mol/ZincTest/zinc_707.mol	0.375000	0.547723	0.600000
../mol/ZincTest/zinc_7.mol	0.200000	0.353553	0.500000
../mol/ZincTest/zinc_370.mol	0.272727	0.433013	0.500000

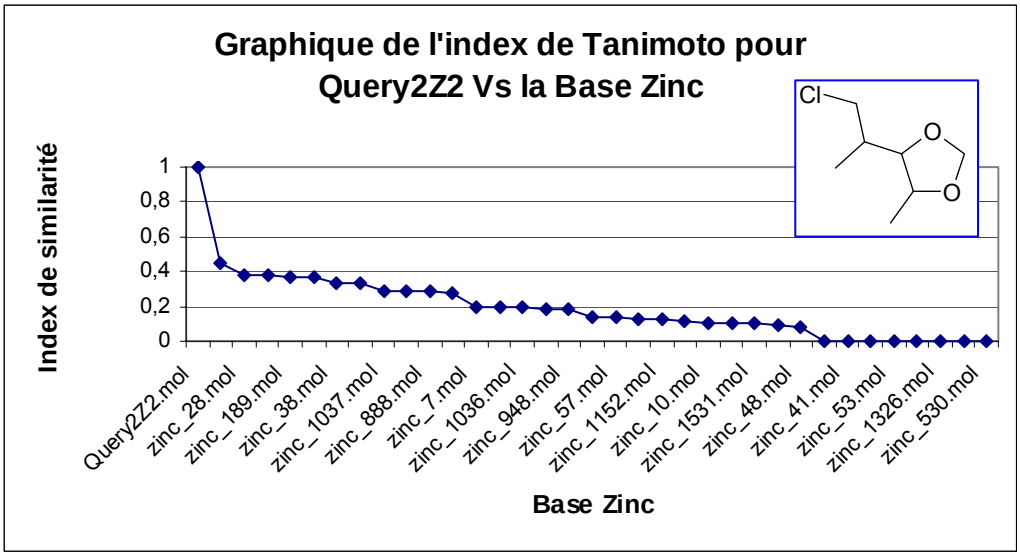
Graphique V.1 Graphique de mesures de similarité pour la base « Zinc » et la cible : Query1Z2.mol

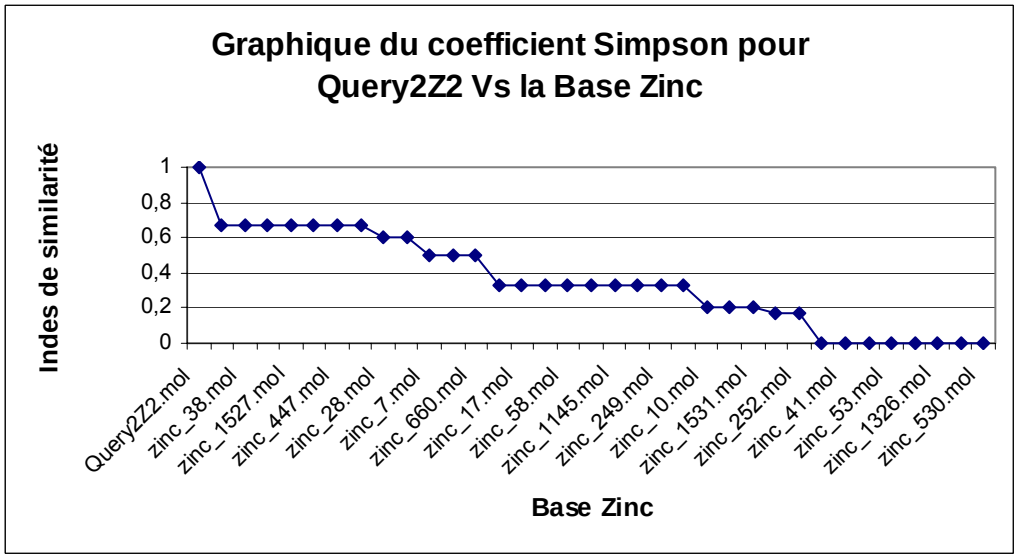


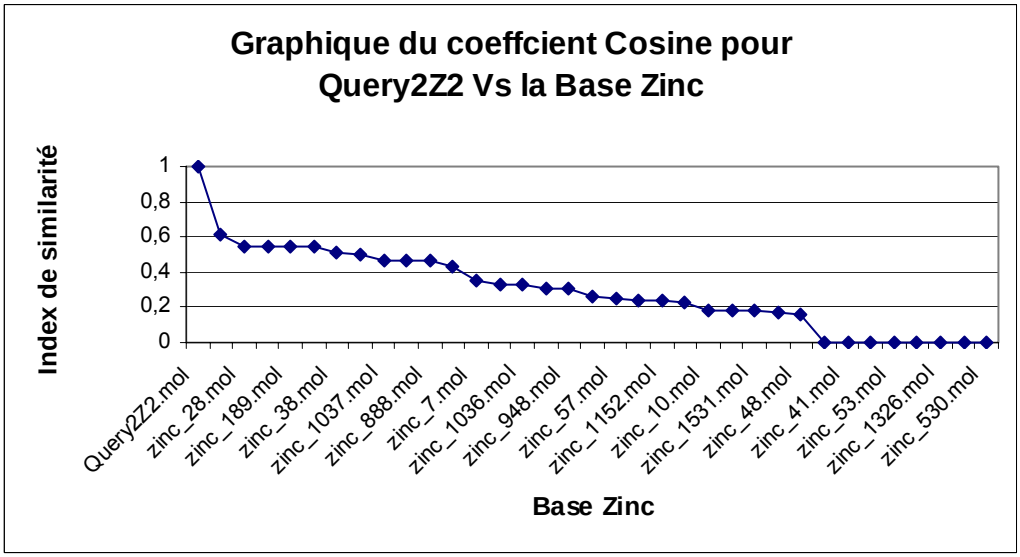




Graphique V.2 Graphique de mesures de similarité pour la base « Zinc » et la cible : Query2Z2.mol







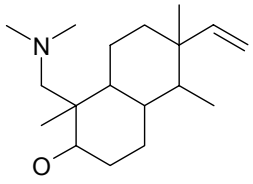


Tableau V.3 Résultats avec cible : Query1Z3.mol

Filename	Tanimoto	Cosinus	Simpson
../mol/ZincQueries/Query1Z3.mol	1.000000	1.000000	1.000000
../mol/ZincTest/zinc_530.mol	0.090909	0.301511	1.000000
../mol/ZincTest/zinc_18.mol	0.500000	0.683763	0.857143
../mol/ZincTest/zinc_48.mol	0.500000	0.683763	0.857143
../mol/ZincTest/zinc_38.mol	0.615385	0.762770	0.800000
../mol/ZincTest/zinc_7.mol	0.533333	0.696311	0.727273
../mol/ZincTest/zinc_948.mol	0.384615	0.569803	0.714286
../mol/ZincTest/zinc_17.mol	0.307692	0.492366	0.666667
../mol/ZincTest/zinc_1036.mol	0.307692	0.492366	0.666667
../mol/ZincTest/zinc_1152.mol	0.166667	0.348155	0.666667
../mol/ZincTest/zinc_189.mol	0.428571	0.603023	0.666667
../mol/ZincTest/zinc_252.mol	0.307692	0.492366	0.666667

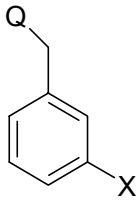
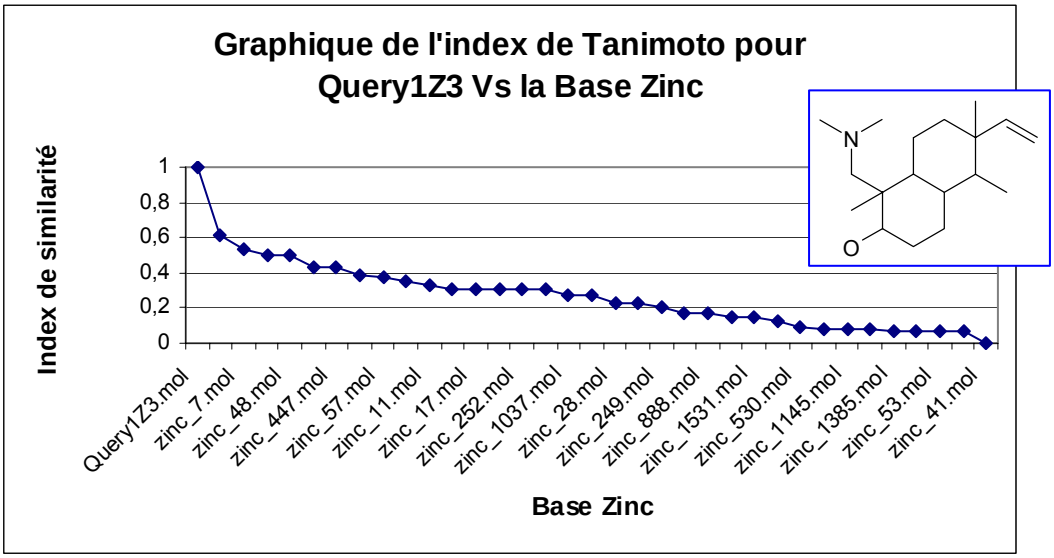
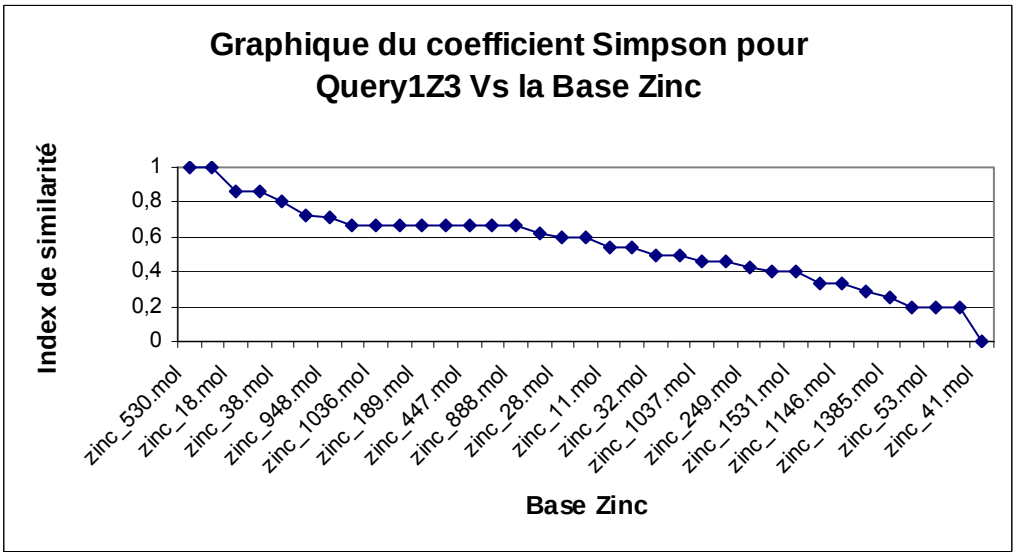


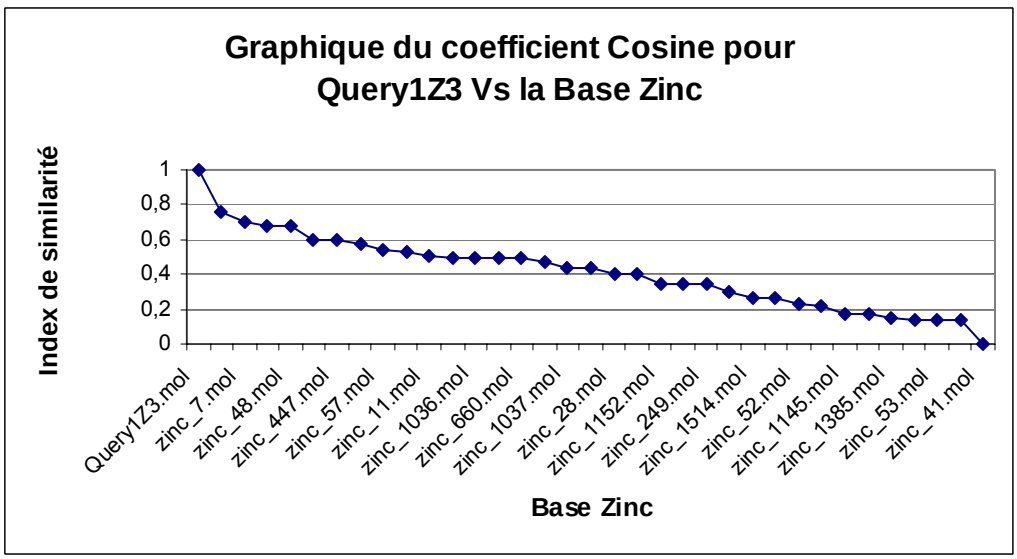
Tableau V.4 Résultats avec cible : Query3Z.mol

Filename	Tanimoto	Cosinus	Simpson
../mol/ZincQueries/Query3Z.mol	1.000000	1.000000	1.000000
../mol/ZincTest/zinc_28.mol	0.400000	0.632456	1.000000
../mol/ZincTest/zinc_249.mol	0.285714	0.534522	1.000000
../mol/ZincTest/zinc_10.mol	0.166667	0.316228	0.500000
../mol/ZincTest/zinc_11.mol	0.071429	0.196116	0.500000
../mol/ZincTest/zinc_17.mol	0.142857	0.288675	0.500000
../mol/ZincTest/zinc_18.mol	0.125000	0.267261	0.500000
../mol/ZincTest/zinc_38.mol	0.090909	0.223607	0.500000
../mol/ZincTest/zinc_48.mol	0.125000	0.267261	0.500000
../mol/ZincTest/zinc_57.mol	0.083333	0.213201	0.500000
../mol/ZincTest/zinc_58.mol	0.090909	0.223607	0.500000
../mol/ZincTest/zinc_7.mol	0.076923	0.204124	0.500000

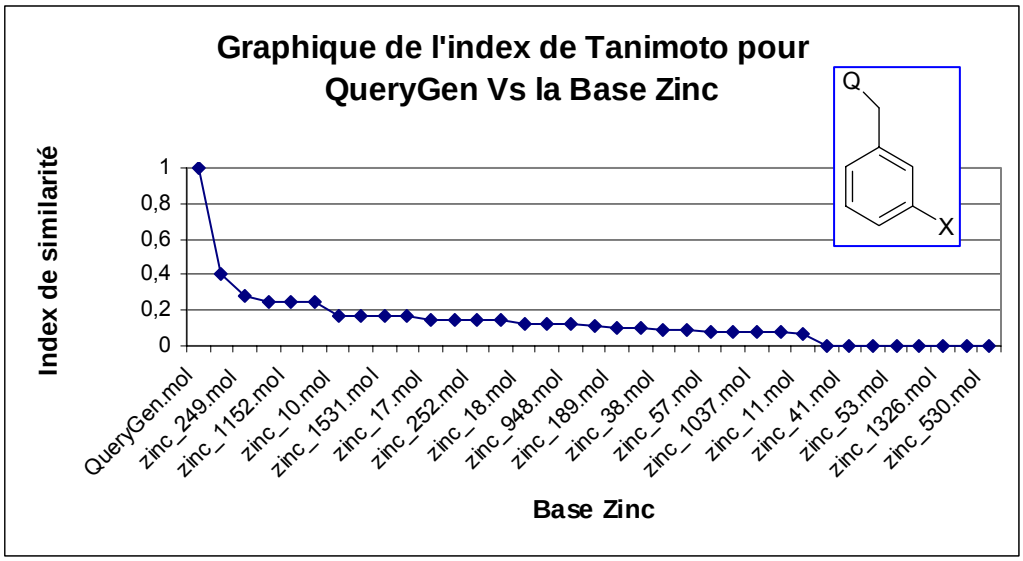
Graphique V.3 Graphique de mesures de similarité pour la base « Zinc » et la cible : Query1Z3.mol

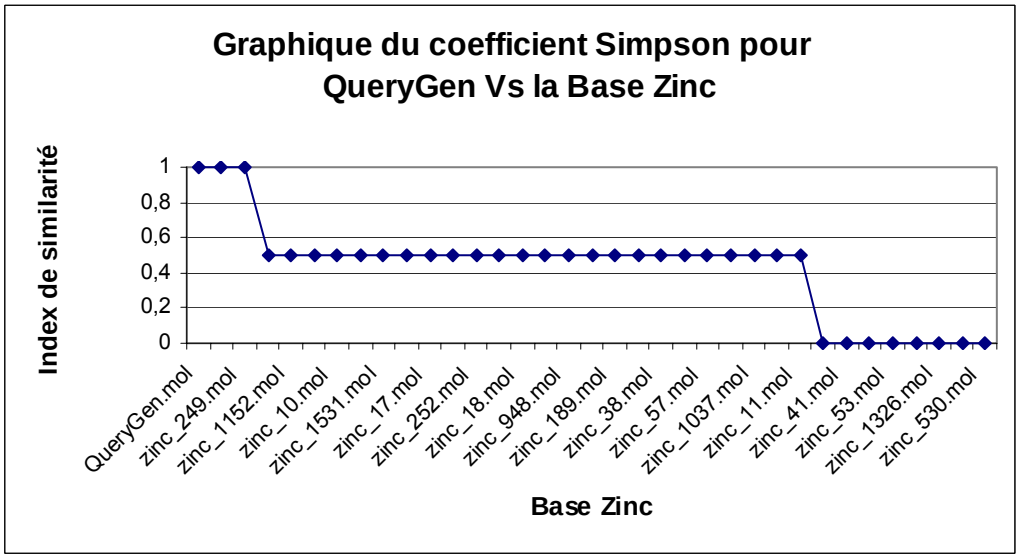


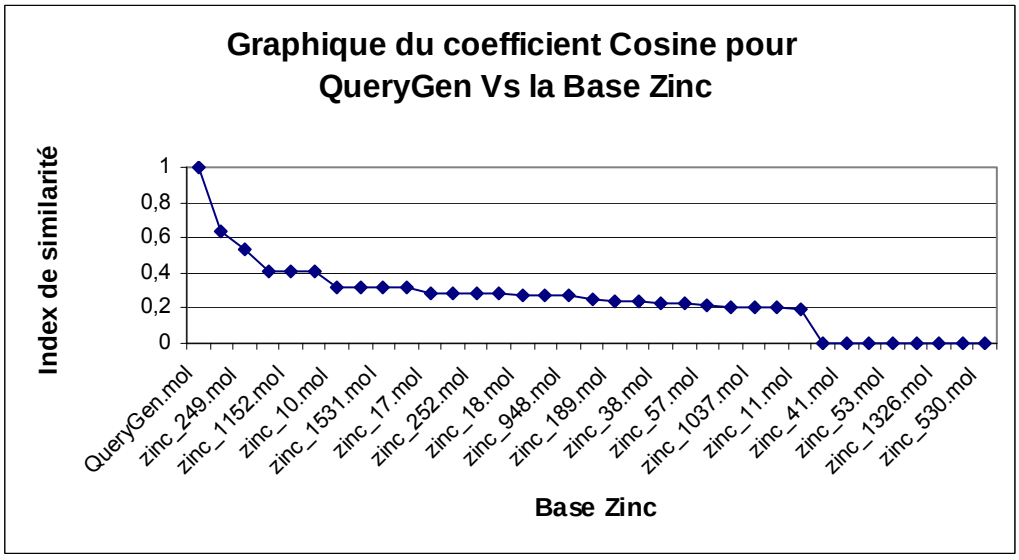




Graphique V.4 Graphique de mesures de similarité pour la base « Zinc » et la cible : Query3Z.mol







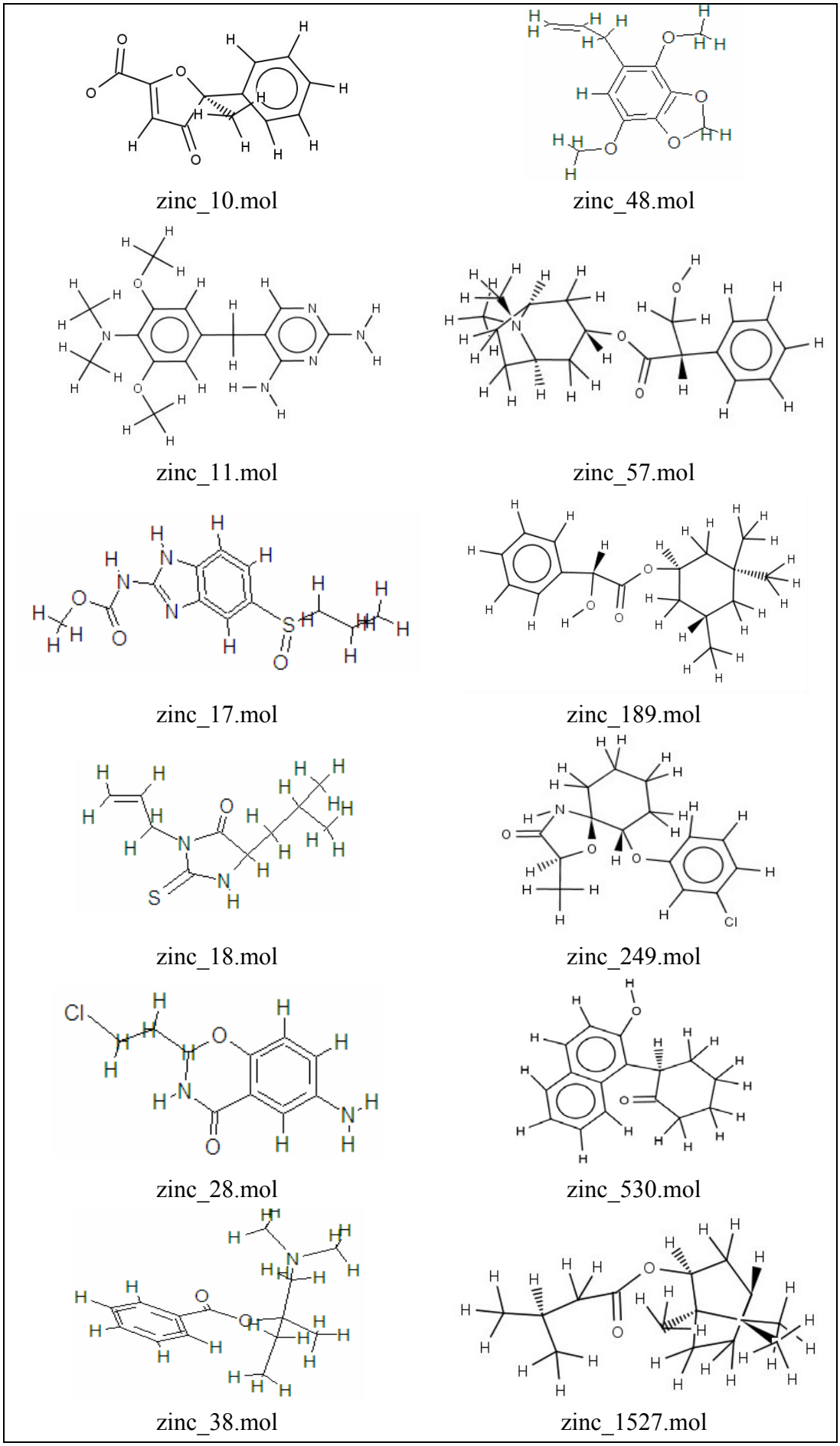


Figure V.1. Quelques molécules de la base « Zinc » appartenant au « top 10 » des tableaux V.1-V.4

Les résultats généraux de l'analyse montrent que pour 100% des molécules, les recherches exactes que MolDiA effectue sont correctes. Par contre, seul l'index de Tanimoto ne produit pas de fausses valeurs unité. On observe également que le calcul de la similarité est différent en utilisant les indices Cosinus ou Simpson. Ceci est montré par le nombre de molécules dont la mesure de similarité est supérieure ou égale à 0,8, trouvées par chaque index (voir tableau ci dessous). L'index de Simpson donne lieu à trois valeurs unités fausses (faux homomorphismes) pour les molécules Query1Z3 et Query3Z. Ceci est dû en partie à la petite taille des molécules (donc peu de fragments dans le vecteur descripteur) et à la présence d'atomes génériques dans Query3Z (recherche de sous-structures). L'inspection de la formule met en évidence que l'index de Simpson ne prend pas en compte l'absence de fragments pour le calcul de la similarité. Pour Query3Z en particulier, le faux homomorphisme trouvé avec l'index de Simpson peut être interprété comment étant en fait un isomorphisme, car la comparaison d'une cible avec des atomes génériques revient à faire une recherche sous-structurale sur la molécule test.

	Mesures de Sim ≥ 0,8							
	Query1Z2		Query2Z2		Query1Z3		Query3Z	
Tanimoto	1	2,94%	1	2,94%	1	2,94%	1	2,94%
Cosinus	2	5,88%	1	2,94%	1	2,94%	1	2,94%
Simpson	6	17,65%	1	2,94%	5	14,7%	3	8,82

Tableau V.5 Nombre et % de molécules avec Is ≥ 0.8 trouvés avec l'outil par rapport aux requêtes.

Les résultats du « *top 10* » pour les mesures de similarité structurale entre la molécule cible et les molécules de la base de Tests, pour les trois indices étudiés (Tanimoto, Simpson et Cosinus) sont indiqués dans les tableaux V.1-V.4. Les molécules en tête du classement sont montrées dans la figure V.1.

Si on effectue maintenant une analyse des résultats cible par cible, on observe que pour la cible Query1Z2, les tendances pour les trois indices sont assez claires. Plus l'indice est restrictif, moins

de correspondances seront trouvées entre les molécules. Ainsi, l'index de Tanimoto (le plus restrictif des indices) trouve une seule molécule avec Is (index de similarité) au-delà de 0.5 : zinc_10. Ceci s'explique par le système aromatique et le groupement acide carboxylique en commun avec la cible. L'application de l'index de Cosinus et de Simpson donne respectivement 7 et 13 molécules avec un Is> 0.5, dont des molécules moins similaires (d'un point de vue structural) à la cible que zinc_10. Cet effet est d'avantage marqué avec des molécules avec peu de fragments représentatifs (Query3Z) ou qui ont des fragments courants (C-C).

Le graphique du coefficient Simpson pour Query3Z présente ainsi des grands paliers de similarité : le premier à un, le deuxième à 0,5 et troisième à zéro. Car seuls trois fragments décrivent la molécule entière. L'usage d'un indice qui prend en compte l'ensemble des données (fragments correspondants et fragments absents) contourne ce problème. Les systèmes cycliques étant très répandus en chimie pharmaceutique, la cible Query1Z3 produit des valeurs de similarité assez élevées pour l'échantillon de la base ZINC étudiée. Finalement, la même raison (la nature et vocation de la base) explique pourquoi la présence de fragments de type (C-X) dans Query2Z2 et Query3Z limite en partie le nombre de correspondances trouvées pour ces deux molécules.

V.1.2 Résultats avec la base « Random »

L'analyse 1-N avec les molécules de la base « Random » nous a permit d'évaluer la capacité de l'outil pour détecter et analyser des sous-structures qui sont moins courantes dans la base « Zinc ». Des exemples de ces fragments sont des sous-structures de nature hydrophobe ou aliphatiques comme ceux contenus dans RandSel100_16.mol, ainsi que quelques systèmes cycliques non aromatiques (RandSel100_29.mol).

Malgré le manque de systèmes cycliques dans «Random» (ce qui n'est pas très représentatif de la diversité moléculaire des grandes bases de molécules actuelles) on observe des bons résultats de détection de ses systèmes. D'autres groupes fonctionnels d'usage courant en chimie ont été testés à travers l'usage de RandSel100_51.mol et de RandSel100_74.mol comme molécules requêtes.

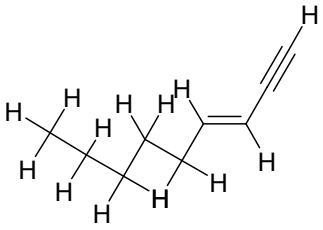


Tableau V.6 Résultats avec cible : RandSel100_16.mol

FileName	Tanimoto	Simpson	Cosinus
../mol/RandSel100/RandSel100_16.mol	1	1	1
../mol/RandSel100/RandSel100_4.mol	0,5	1	0,70
../mol/RandSel100/RandSel100_52.mol	0,33	1	0,57
../mol/RandSel100/RandSel100_6.mol	0,71	0,83	0,83
../mol/RandSel100/RandSel100_35.mol	0,71	0,83	0,83
../mol/RandSel100/RandSel100_46.mol	0,71	0,83	0,83
../mol/RandSel100/RandSel100_89.mol	0,62	0,83	0,77
../mol/RandSel100/RandSel100_92.mol	0,62	0,83	0,77
../mol/RandSel100/RandSel100_1.mol	0,57	0,8	0,73
../mol/RandSel100/RandSel100_15.mol	0,57	0,8	0,73
../mol/RandSel100/RandSel100_29.mol	0,57	0,8	0,73

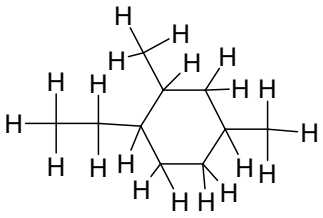
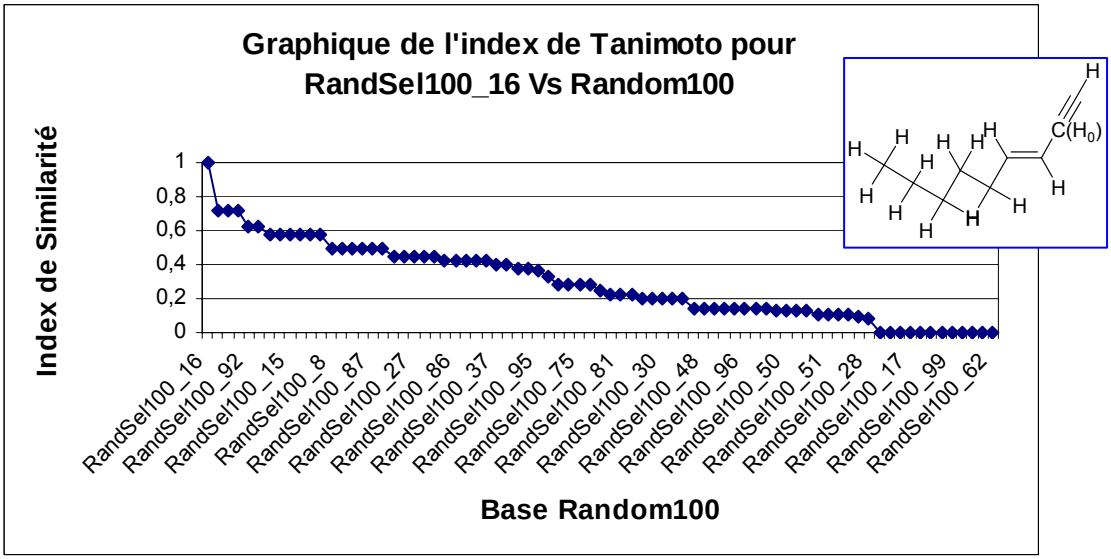
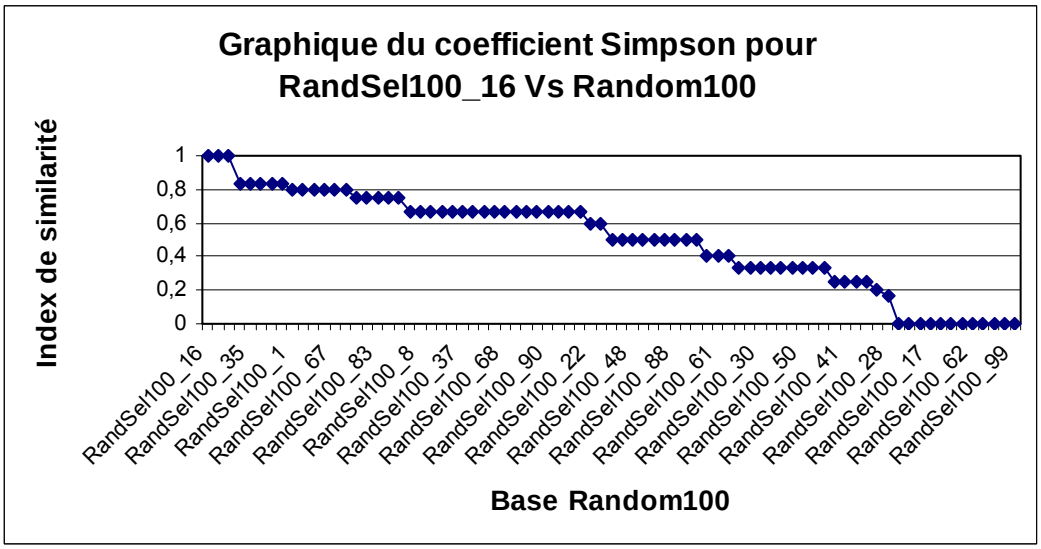


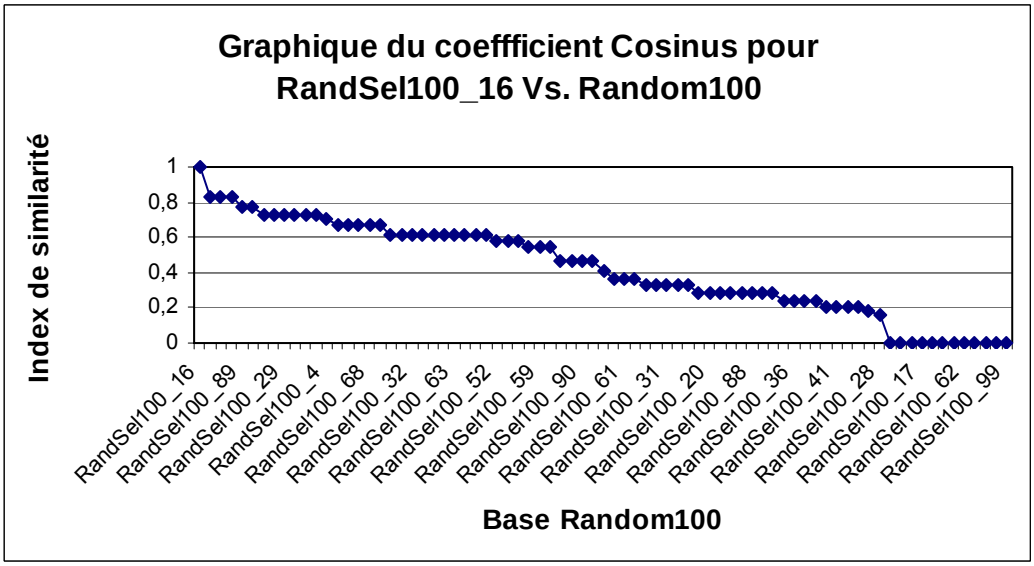
Tableau V.7 Résultats avec cible : RandSel100_29.mol

FileName	Tanimoto	Simpson	Cosinus
../mol/RandSel100/RandSel100_29.mol	1	1	1
../mol/RandSel100/RandSel100_15.mol	0,66	0,8	0,8
../mol/RandSel100/RandSel100_16.mol	0,57	0,8	0,73
../mol/RandSel100/RandSel100_2.mol	0,5	0,8	0,67
../mol/RandSel100/RandSel100_6.mol	0,57	0,8	0,73
../mol/RandSel100/RandSel100_8.mol	0,57	0,8	0,73
../mol/RandSel100/RandSel100_27.mol	0,5	0,8	0,67
../mol/RandSel100/RandSel100_32.mol	0,5	0,8	0,67
../mol/RandSel100/RandSel100_35.mol	0,57	0,8	0,73
../mol/RandSel100/RandSel100_37.mol	0,44	0,8	0,63
../mol/RandSel100/RandSel100_40.mol	0,66	0,8	0,8

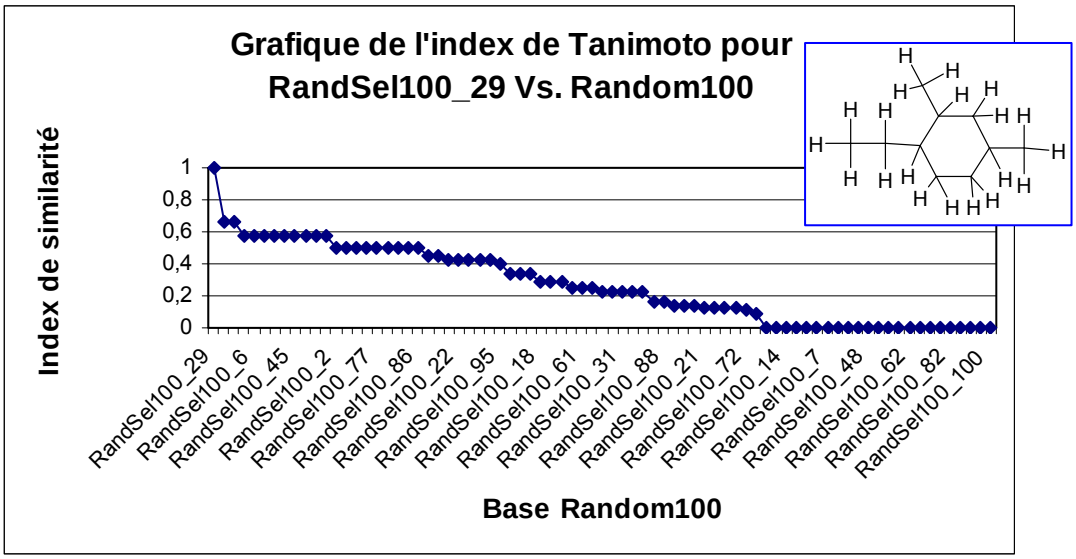
Graphique V.5 Graphique de mesures de similarité pour « Random » et RandSel100_16.mol

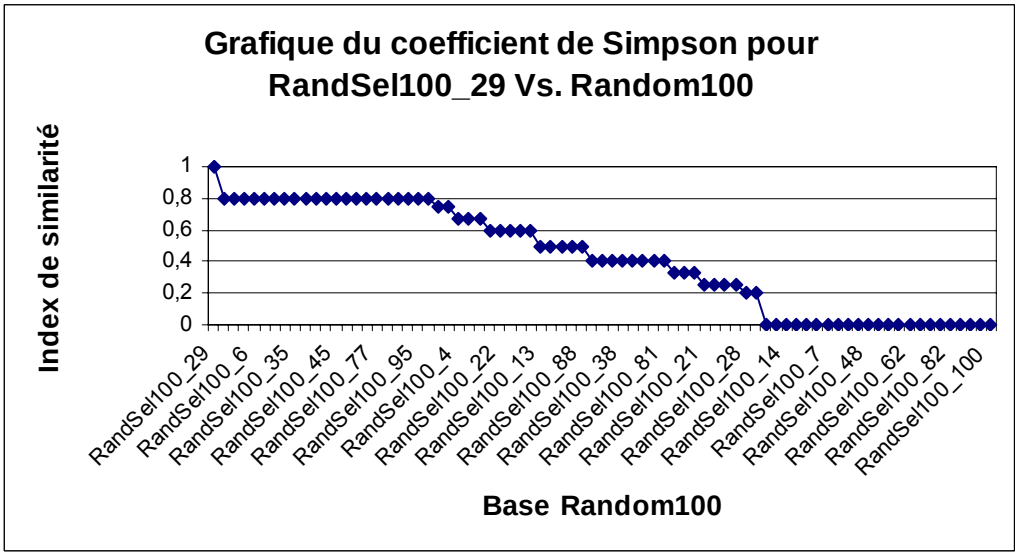


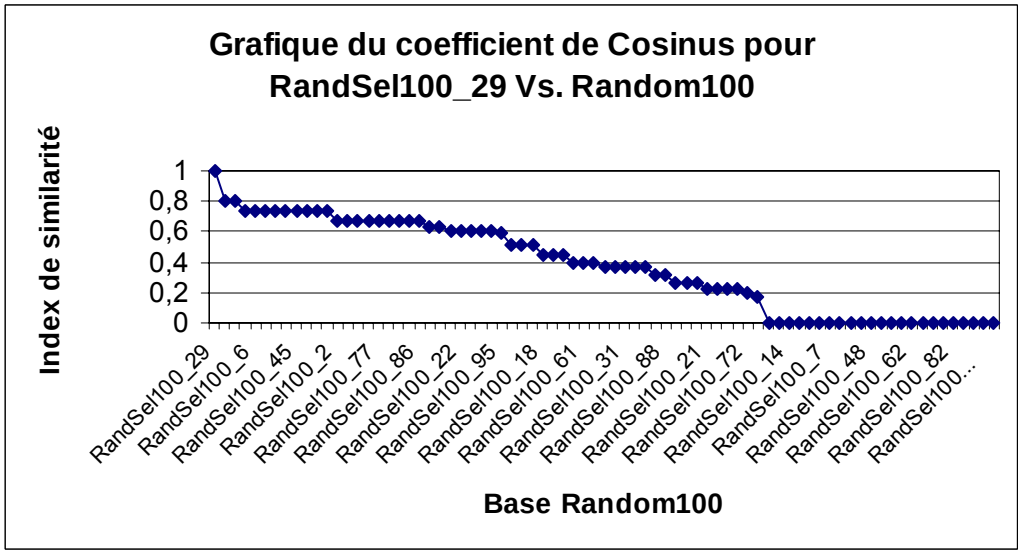




Graphique V.6 Graphique de mesures de similarité pour « Random » et RandSel100_29.mol







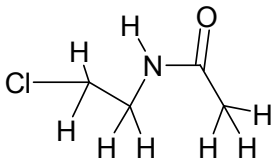


Tableau V.8 Résultats avec cible : RandSel100_51.mol

FileName	Tanimoto	Simpson	Cosinus
../mol/RandSel100/RandSel100_51.mol	1	1	1
../mol/RandSel100/RandSel100_41.mol	1	1	1
../mol/RandSel100/RandSel100_13.mol	0,5	1	0,70
../mol/RandSel100/RandSel100_88.mol	0,5	1	0,70
../mol/RandSel100/RandSel100_24.mol	0,42	0,75	0,61
../mol/RandSel100/RandSel100_27.mol	0,37	0,75	0,56
../mol/RandSel100/RandSel100_28.mol	0,37	0,75	0,56
../mol/RandSel100/RandSel100_31.mol	0,42	0,75	0,61
../mol/RandSel100/RandSel100_43.mol	0,5	0,75	0,67
../mol/RandSel100/RandSel100_76.mol	0,42	0,75	0,61
../mol/RandSel100/RandSel100_81.mol	0,5	0,75	0,67

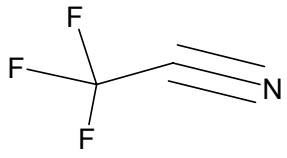
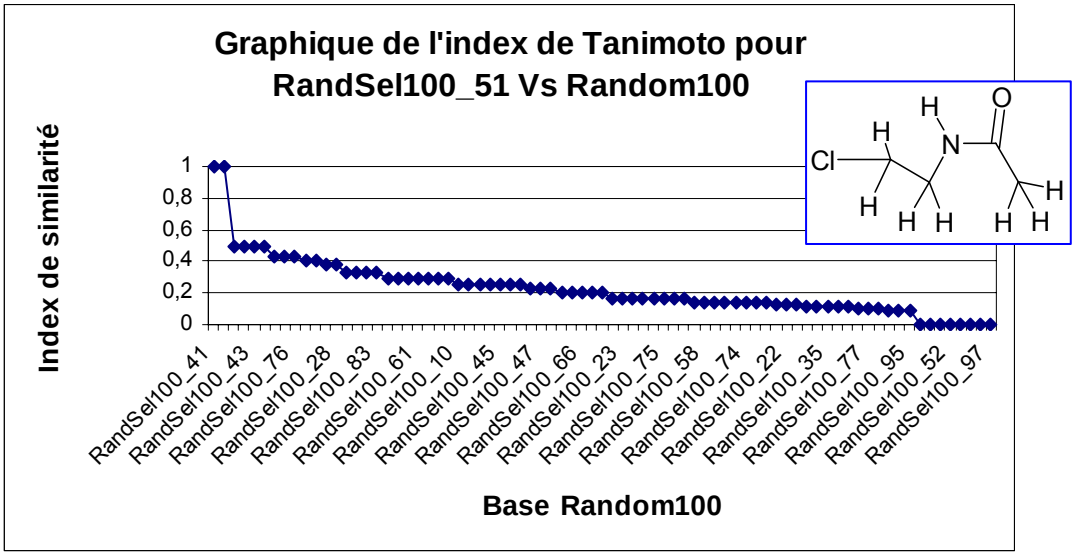
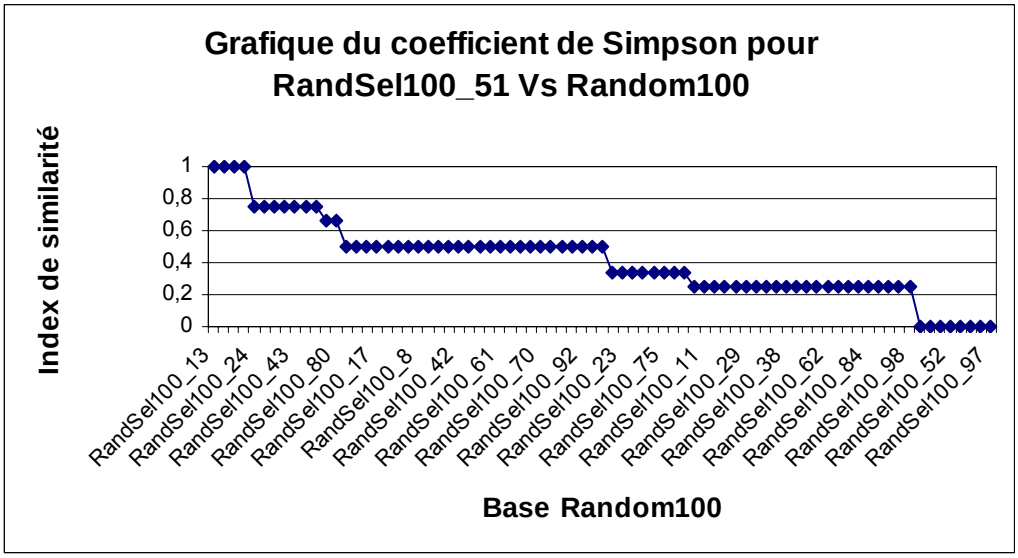


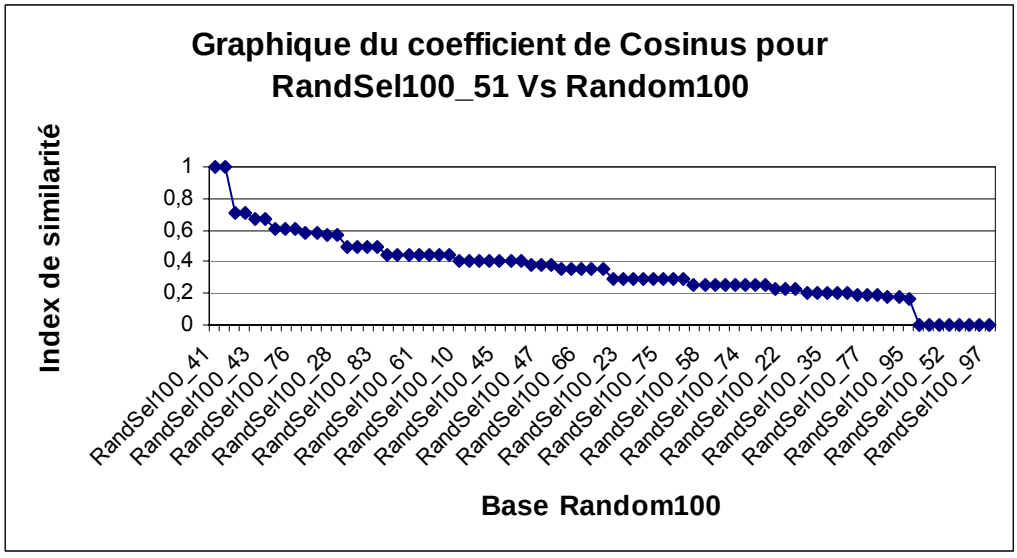
Tableau V.9 Résultats avec cible : RandSel100_74.mol

FileName	Tanimoto	Simpson	Cosinus
../mol/RandSel100/new-RandSel100_74.mol	1	1	1
../mol/RandSel100/RandSel100_14.mol	0,5	1	0,70
../mol/RandSel100/RandSel100_60.mol	0,6	0,75	0,75
../mol/RandSel100/RandSel100_13.mol	0,2	0,5	0,35
../mol/RandSel100/RandSel100_19.mol	0,2	0,5	0,35
../mol/RandSel100/RandSel100_76.mol	0,25	0,5	0,40
../mol/RandSel100/RandSel100_88.mol	0,2	0,5	0,35
../mol/RandSel100/RandSel100_26.mol	0,16	0,33	0,28
../mol/RandSel100/RandSel100_10.mol	0,11	0,25	0,20
../mol/RandSel100/RandSel100_15.mol	0,12	0,25	0,22
../mol/RandSel100/RandSel100_17.mol	0,14	0,25	0,25

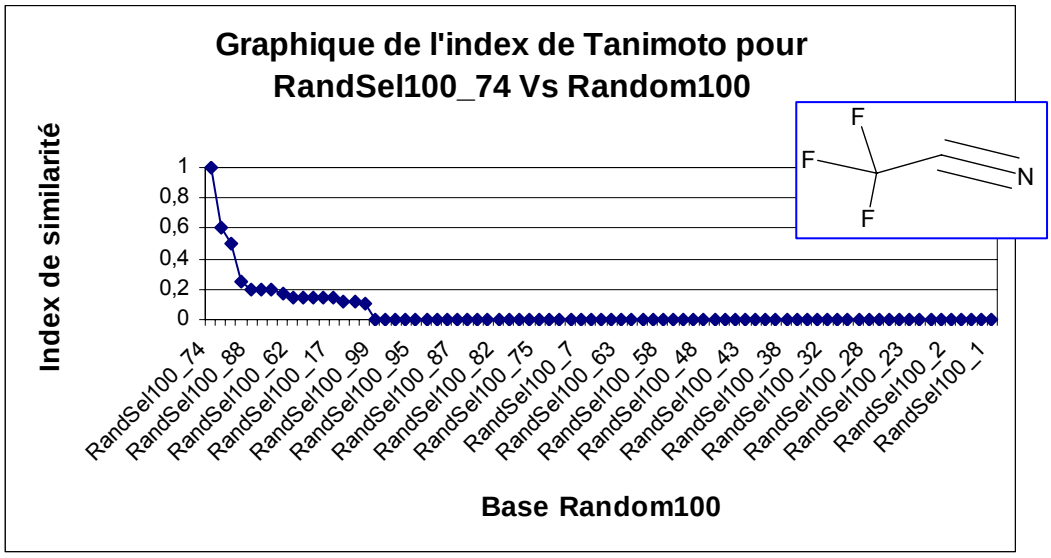
Graphique V.7 Graphique de mesures de similarité pour « Random » et RandSel100_51.mol

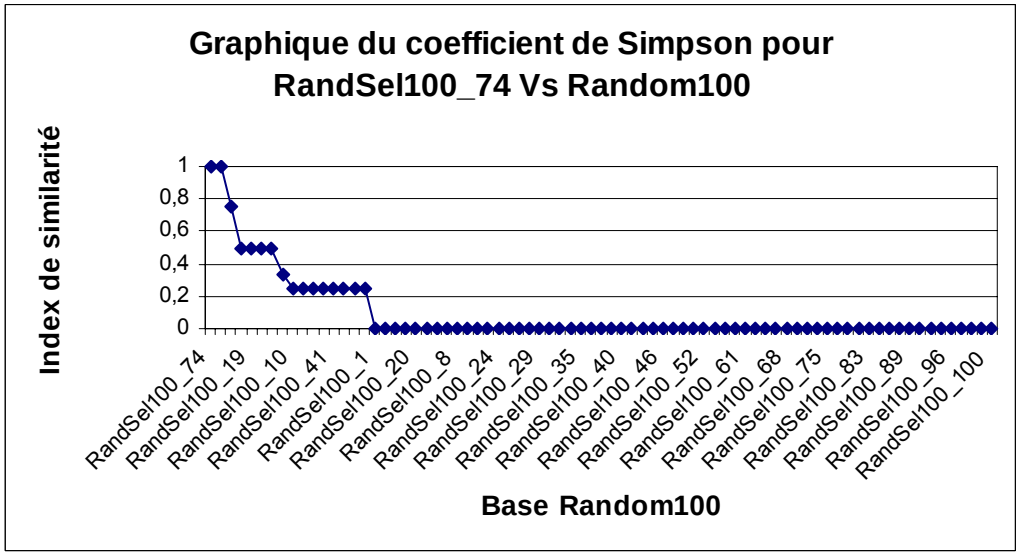


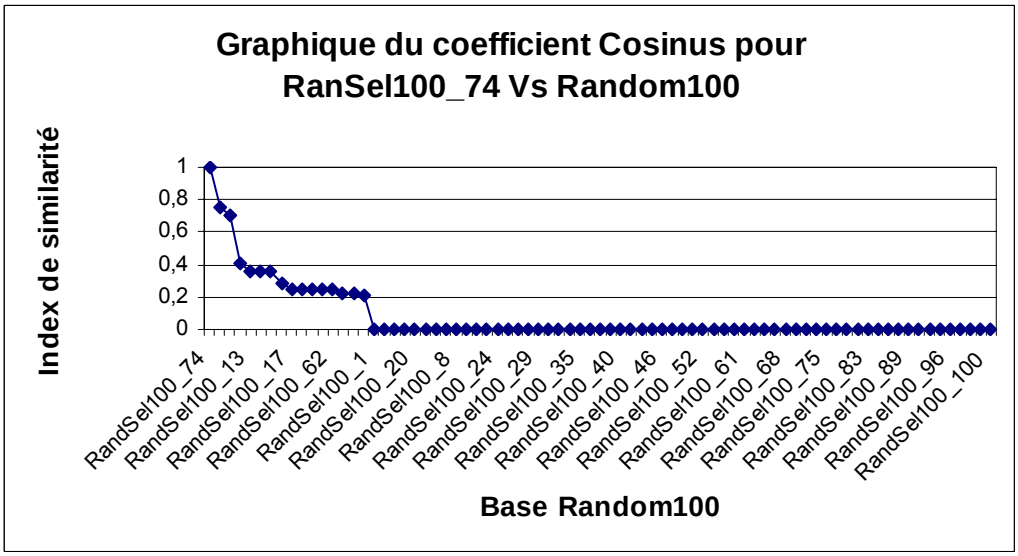


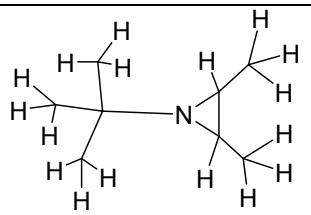


Graphique V.8 Graphique de mesures de similarité pour « Random » et RandSel100_74.mol

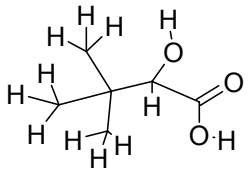




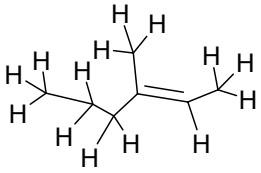




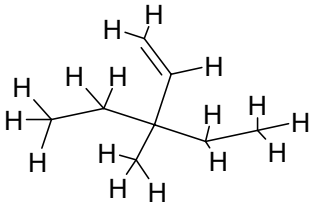
RandSel100_2.mol



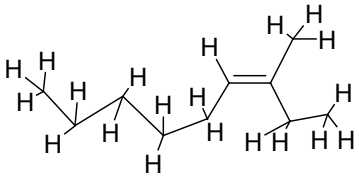
RandSel100_27.mol



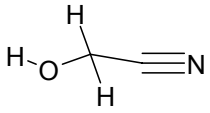
RandSel100_4.mol



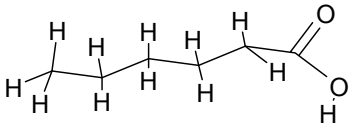
RandSel100_35.mol



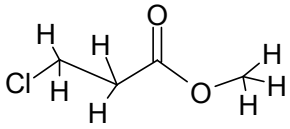
RandSel100_6.mol



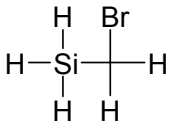
RandSel100_19.mol



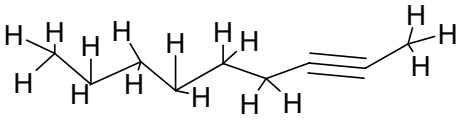
RandSel100_8.mol



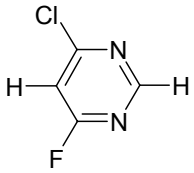
RandSel100_41.mol



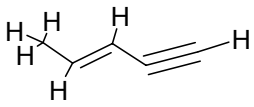
RandSel100_13.mol



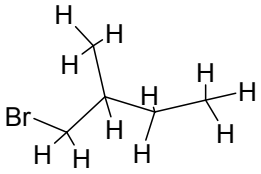
RandSel100_46.mol



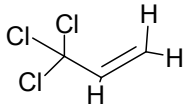
RandSel100_14.mol



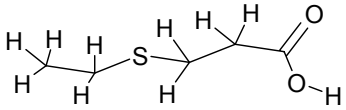
RandSel100_52.mol



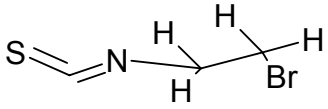
RandSel100_15.mol



RandSel100_60.mol



RandSel100_24.mol



RandSel100_88.mol

Figure V.2. Quelques molécules de la base « Random » appartenant au « top 10 »

Après inspection des tableaux V.5-V.9, on remarque que 100% des requêtes ont abouti à des résultats de recherche d'homomorphismes. Toutefois, le nombre de mesures varie d'index en index et en fonction de la molécule requête. Plus la cible est petite, moins de correspondances avec un haut indice de similarité seront trouvées. Par contre, le nombre de mesures de similarité entre 0,3 et 0,6 augmente. Ceci est particulièrement valable en utilisant l'index de Simpson dans le calcul. La cible qui présente le moins de correspondances est RandSel100_74, où en moyenne, 79% des molécules présentent une mesure de similarité égale à zéro envers cette requête. De faux isomorphismes ont été trouvés avec tous les indices. Une analyse plus détaillée du phénomène de faux isomorphismes sera donnée dans la section V.3.2.

	Mesures de Sim ≥ 0,8							
	RandSel100_16		RandSel100_29		RandSel100_51		RandSel100_74	
Tanimoto	1	1,3%	1	1,3%	2	2,6%	1	1,3%
Cosinus	4	5,19%	3	3,9%	2	2,6%	1	1,3%
Simpson	14	18,18%	22	28,57%	4	5,19%	2	2,6%

Tableau V.10 Nombre et % de molécules avec Is ≥ 0.8, trouvés avec MolDiA par rapport aux quatre molécules requêtes. Quelques molécules test composant la base « Random » sont montrés dans la figure V.2

La composition de la base « Random » est élucidée après examen du graphique V.5 : 13 molécules (Tanimoto), 48 molécules (Simpson) et 43 molécules (Cosinus) ont des Is> 0.5, ceci nous permet d'affirmer que la base de test est composée principalement des molécules ressemblant à RandSel100_16. La base est donc homogène et contient principalement des molécules petites, la plupart avec une grande concentration des sous-structures aliphatiques. Ces caractéristiques se traduisent en une pauvre diversité moléculaire à l'égard, par exemple, d'une base d'origine pharmaceutique comme celle du « Zinc ». Malgré tout, « Random » comprend également des fragments cycliques, mais en plus petite quantité que la base « Zinc ». D'autres tests (non montrés ici), mettent en évidence la quasi absence des cycles aromatiques et de systèmes multi cycliques.

Les sous-structures de type « groupe fonctionnel » sont également moins courantes comme en témoignent les graphiques V.7 et V.8. La base est particulièrement pauvre en sous-structures contenant des atomes de Fluor (tous indices confondus), comme en témoigne le graphique V.8 qui montre les mesures de similarité entre la base « Random » et RandSel100_74.

V.1.3 Comparaison des indices selon le rang

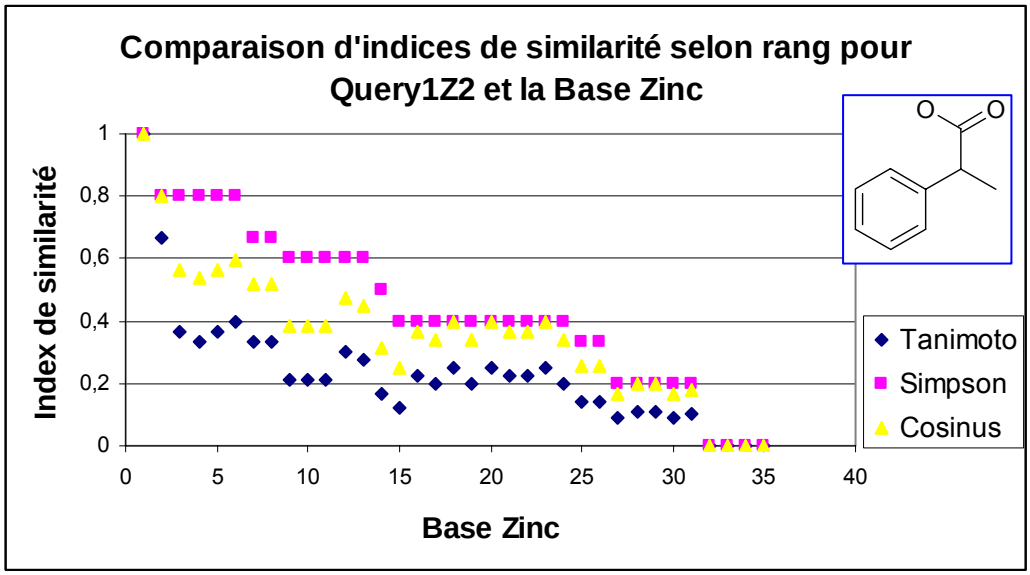
Pour étudier le comportement des mesures de similarité/diversité utilisées, à savoir, Tanimoto, Simpson et Cosinus, nous pouvons effectuer un graphique « indice de similarité vs noms de molécules » dans lequel on peut superposer les résultats pour les 3 indices. Les tendances de trois mesures de similarité peuvent être ainsi mieux appréciées en faisant un classement des molécules de la base utilisée. Cette étude a été réalisée pour les molécules des deux bases de test.

V.1.3.1 Graphiques de comparaison d'indices avec la base « Zinc »

Dans les figures suivantes, on constate que la tendance des graphiques est monotone pour les trois indices. Ceci montre que tous les indices ont indiqué comme similaires ou dissimilaires à peu près les mêmes molécules. Toutefois, dans le graphique comparatif avec Query1Z3, il y a des molécules où la mesure de similarité en utilisant l'index de Simpson contredit les résultats des autres indices. Ces « pics » qui rompent la monotonie du graphique, sont probablement dus à une mauvaise description de la molécule test, ce qui entraîne une faible capacité de comparaison avec la cible. Ces trois graphiques sont assez différents de ceux obtenus avec Query3Z. Ceci met en évidence que l'usage des atomes génériques, peut appauvrir d'une part l'information des correspondances (car il y a plus de possibilités que si l'on n'utilise pas les atomes génériques) mais permet également d'étudier d'une manière rapide la présence ou l'absence d'un groupe fonctionnel ou d'une sous-structure particulière dans la base étudiée.

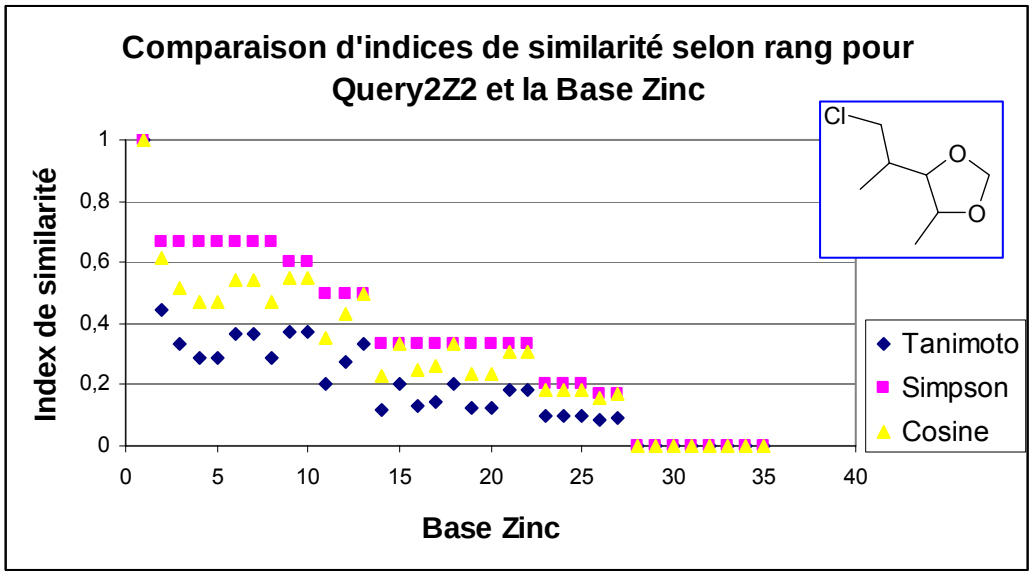
Graphique V.9 Comparaison de 3 mesures de similarité pour la base « Zinc » et la cible :

Query1Z2.mol



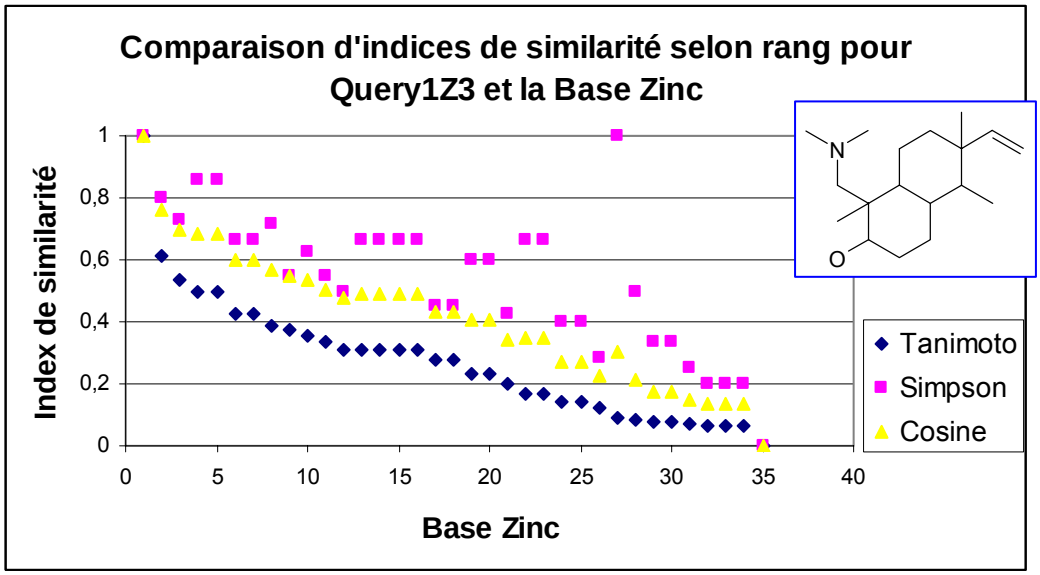
Graphique V.10 Comparaison de 3 mesures de similarité pour la base « Zinc » et la cible :

Query2Z2.mol



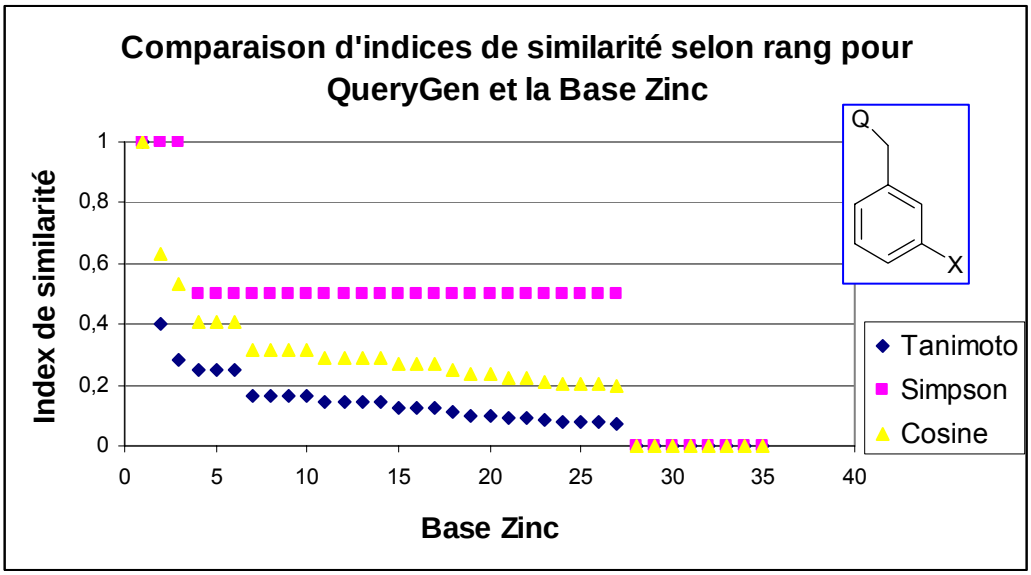
Graphique V.11 Comparaison de 3 mesures de similarité pour la base « Zinc » et la cible :

Query1Z3.mol



Graphique V.12 Comparaison de 3 mesures de similarité pour la base « Zinc » et la cible :

Query3Z.mol

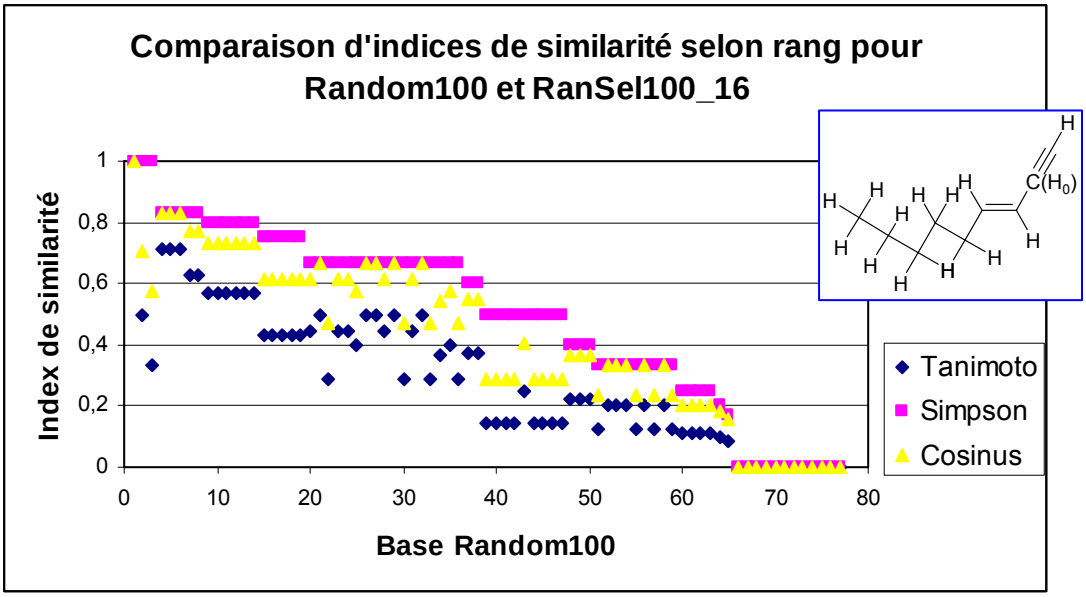


V.1.3.2 Graphiques de comparaison d'indices avec la base « Random »

La comparaison d'indices de similarité pour RanSel100_16, RanSel100_29 et RanSel100_51 montre un comportement monotone, interrompu par quelques molécules pour lesquels les 3 mesures de similarité ne se correspondent pas. On observe également un décalage d'environ 0,1-0,2 entre chaque index. Ce décalage diminue vers les valeurs extrêmes (zéro et un). Ceci est dû à la formule de calcul des indices et à la normalisation qui succède.

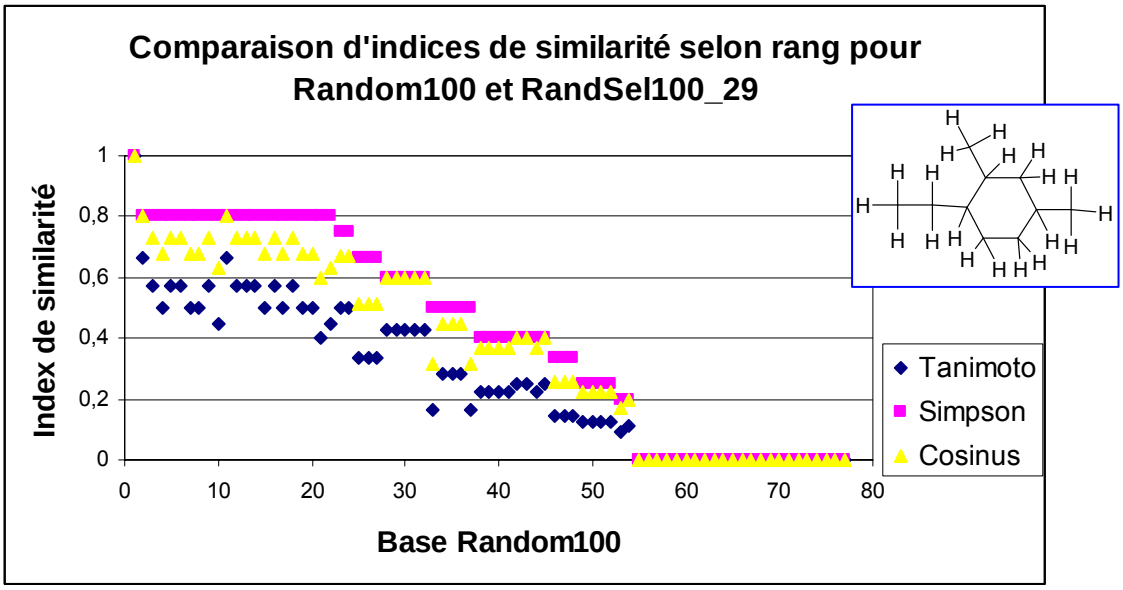
Un comportement différent est observé pour le graphique de comparaison d'indices de similarité de RanSel100_74. Une grande partie des mesures de similarité est nulle, car la molécule requête contient des sous-structures qui sont rares dans la base Random, notamment des fragments du type R-X. Les trois indices ont eu, en moyenne, le même comportement vis-à-vis de cette molécule cible.

Graphique V.13 Comparaison de 3 mesures de similarité pour la base « Random » et la cible : RandSel100_16.mol



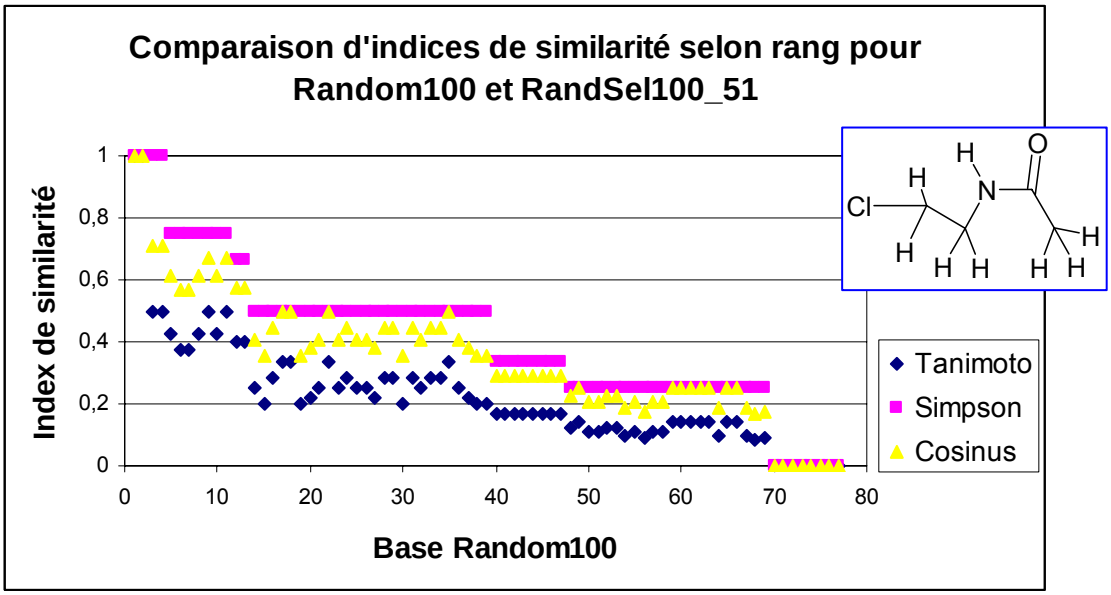
Graphique V.14 Comparaison de 3 mesures de similarité pour la base « Random » et la cible :

RandSel100_29.mol

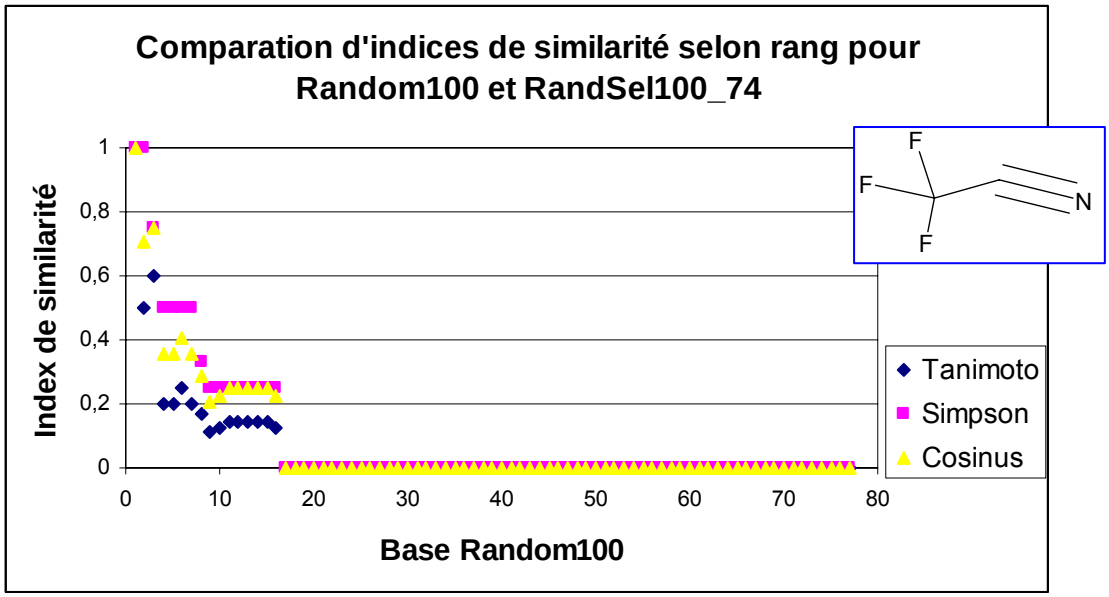


Graphique V.15 Comparaison de 3 mesures de similarité pour la base « Random » et la cible :

RandSel100_51.mol



Graphique V.16 Comparaison de 3 mesures de similarité pour la base « Random » et la cible : RandSel100_74.mol



Pourquoi avoir effectué des études de comparaison de rangs pour des bases moléculaires en utilisant différents critères d'analyse ? La réponse est donnée dans une étude récente:

« La fusion des classements dans les mesures de similarité est généralement plus efficace (en termes de recherche des molécules bio-actives) que des classements basés sur un seul coefficient, sous condition qu'une combinaison appropriée des coefficients soit choisie pour la fusion »

Cette affirmation a été l'une des conclusions d'un travail récent [Holliday2002] sur le comportement de plusieurs mesures de similarité et des résultats obtenus par combinaison de ces mesures. L'étude a été effectuée sur des grandes bases de données chimiques avec un intérêt certain sur la possibilité de retrouver des molécules actives à travers des analyses de similarité.

Mais la « fusion des classements » n'est pas une opération triviale. Elle peut être définie comme la recherche d'une partition optimale de *n* individus lorsque l'on a en entrée *p* partitions différentes de

ces *n* individus. Dans notre cas, 1 individu = 1 molécule, et 1 classement = 1 partition (ce qui équivaut à classer, ou partitionner, les *n* individus en *K* classes).

Ainsi, cette observation fournit une façon simple d'augmenter la performance des systèmes existants pour la recherche de similitude à partir d'analyses sous structurales. Les systèmes actuels comparent les descripteurs d'une structure cible aux descripteurs de chacune des molécules tests dans une base de données. Ceci est fait en utilisant les sous-structures communes et non communes de chaque comparaison pour calculer une mesure de similarité (couramment le coefficient de Tanimoto). Les résultats obtenus par [Holliday2002] suggèrent que si ces sous-structures communes sont en plus employées pour calculer les valeurs d'autres coefficients (comme cela est proposé dans MolDiA), alors le classement résultant aura un plus grand nombre de composés actifs dans les rangs élevés que si seul le coefficient de Tanimoto est employé.

Ces valeurs additionnelles de coefficient peuvent être calculées à un coût informatique négligeable (puisque les comparaisons de sous-structures ont été déjà effectuées pour le calcul de Tanimoto). Ainsi, l'utilisation de la « fusion de données » ou *data fusion* a comme conséquence une augmentation de l'efficacité de recherche avec seulement une diminution très légère du temps de recherche. Une étude de type « fusion de données » avec les bases moléculaires utilisées dans le présent manuscrit et en utilisant toutes les possibilités de calcul proposés dans l'outil, présente un intérêt certain. Des résultats préliminaires ont démontré l'efficacité de cette approche mais n'ont pas été inclus dans le présent manuscrit.

V.1.4 Comparaison des indices selon la complexité

Une autre caractéristique qui peut être intéressante à étudier, est le comportement des indices utilisés (Tanimoto, Sympson et Cosinus) selon la taille des molécules appartenant aux bases de test. Pour cela, il suffit de calculer le nombre d'atomes de chaque molécule et de construire un graphique d'indice de similarité versus le nombre d'atomes dans les molécules test. Les « nuages » de points serviront d'indices pour déterminer la densité de la population moléculaire. Les deux bases utilisées

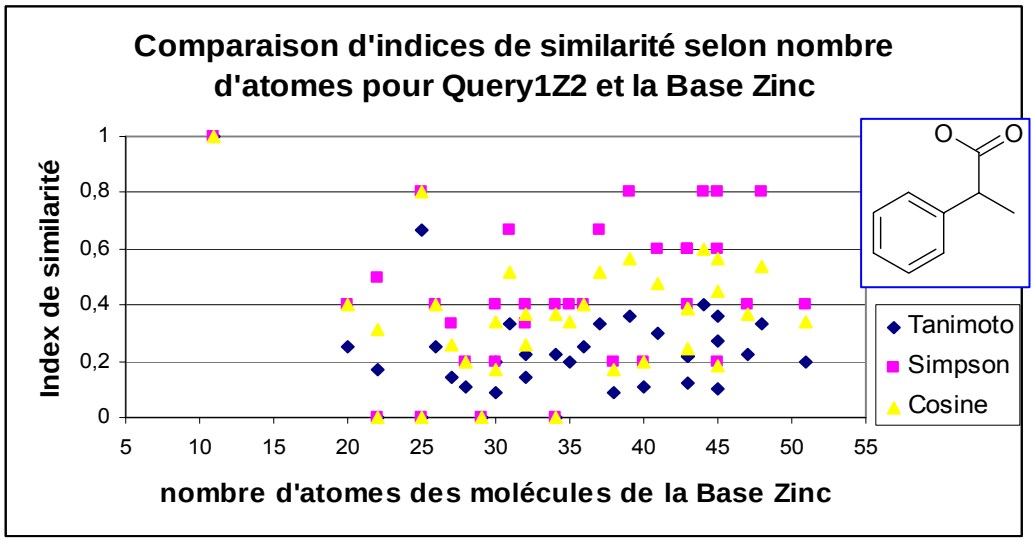
(Zinc et Random) ont comme avantage de contenir des molécules de tailles très différentes, ce qui facilite la comparaison.

V.1.4.1 Graphiques comparatif d'indices avec la base « Zinc »

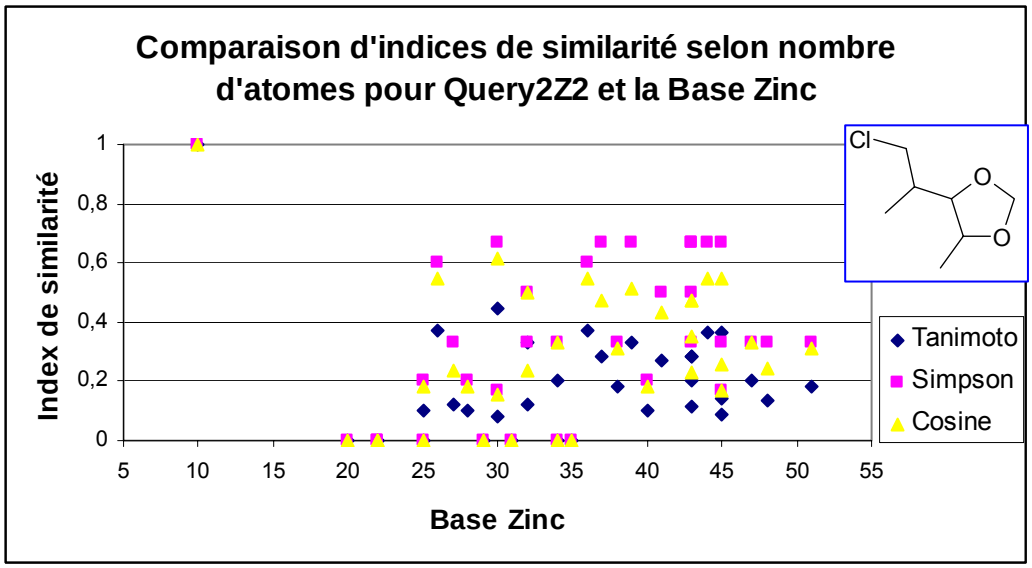
Dans les graphiques V.17-V.20 on observe un nombre élevé de mesures correspondantes aux molécules avec un nombre d'atomes compris entre 20 et 45. On peut déduire que la base « Zinc » est composée des molécules assez grandes, ceci en partie par la nature des molécules appartenant à la base : drogues diverses et molécules d'intérêt pharmaceutique.

Pour la molécule Query1Z2, le « nuage » présente une densité maximale vers 30-45 atomes, avec une mesure de similarité autour de 0,2-0,4. Ceci est logique, puisque la molécule cible est une molécule contenant un système cyclique, et presque toutes les molécules pour lesquelles on a observé une présence de correspondance avaient un ou plusieurs systèmes cycliques, donc un nombre d'atomes supérieur à la requête. En utilisant la cible Query2Z2, la concentration maximale de points est vers 25-45 atomes, avec une mesure de similarité qui varie entre 0,2 et 0,6 tous indices confondus. Là encore, la présence d'un système hétérocyclique dans la cible, conduit à des correspondances avec des systèmes bi et tri-cycliques comportant un nombre d'atomes très supérieur à la cible originale. D'autre part, une assez grande dispersion des valeurs de similarité est obtenue avec Query1Z3. La composition de la molécule ainsi que la présence des nombreux groupes fonctionnels augment les possibilités de correspondances avec les molécules de la base « Zinc ». L'usage d'atomes génériques diminue la reconnaissance « fine » des sous-structures, mais augmente les possibilités de correspondances (malgré une faible mesure de similarité d'environ 0,1-0,5).

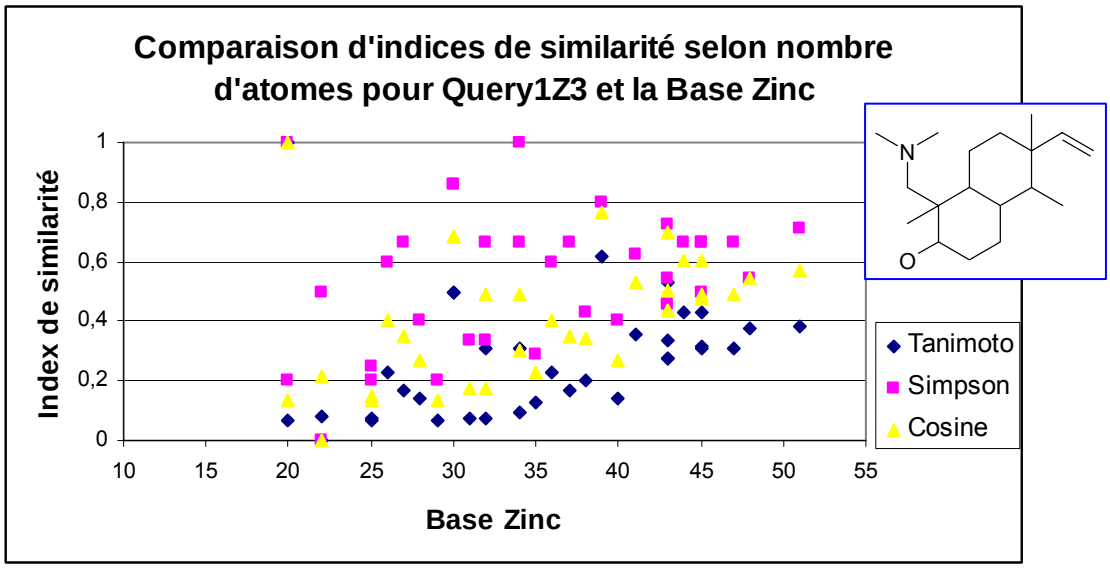
Graphique V.17 Comparaison de trois mesures de similarité en fonction de la taille de molécules de la base « Zinc » pour la cible Query1Z2.mol



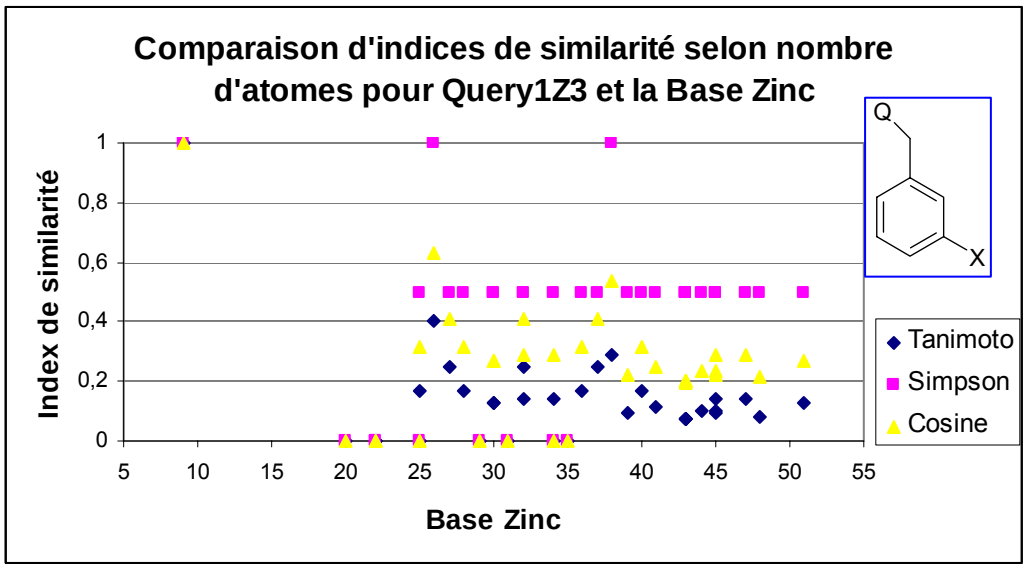
Graphique V.18 Comparaison de trois mesures de similarité en fonction de la taille de molécules de la base « Zinc » pour la cible Query2Z2.mol



Graphique V.19 Comparaison de trois mesures de similarité en fonction de la taille de molécules de la base « Zinc » pour la cible Query1Z3.mol



Graphique V.20 Comparaison de trois mesures de similarité en fonction de la taille de molécules de la base « Zinc » pour la cible Query3Z.mol

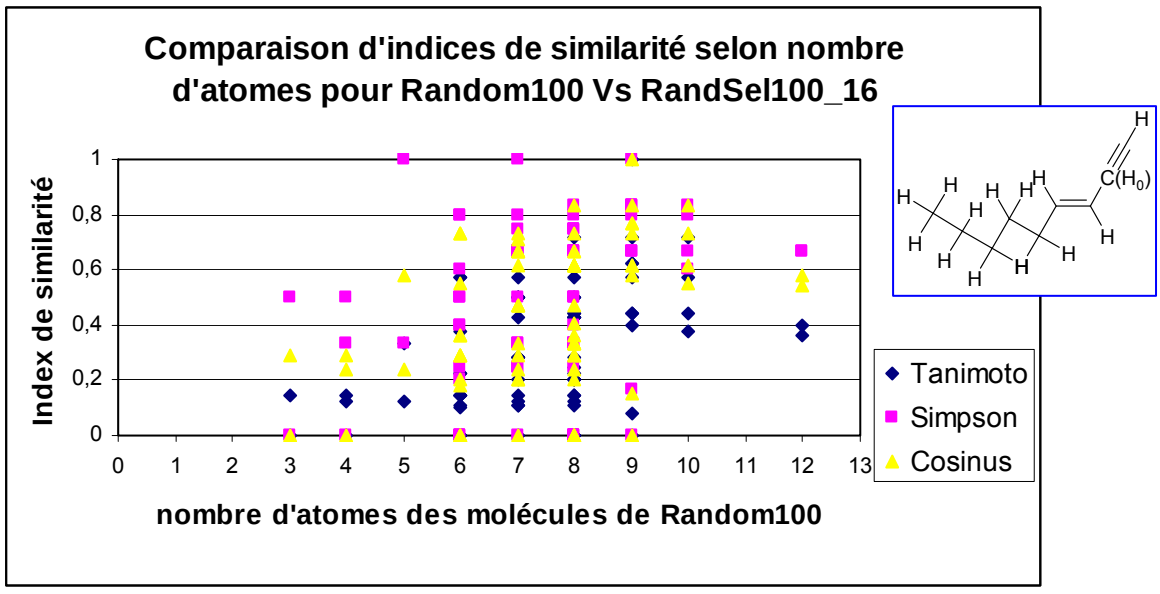


V.1.4.2 Graphiques comparatif d'indices avec la base « Random »

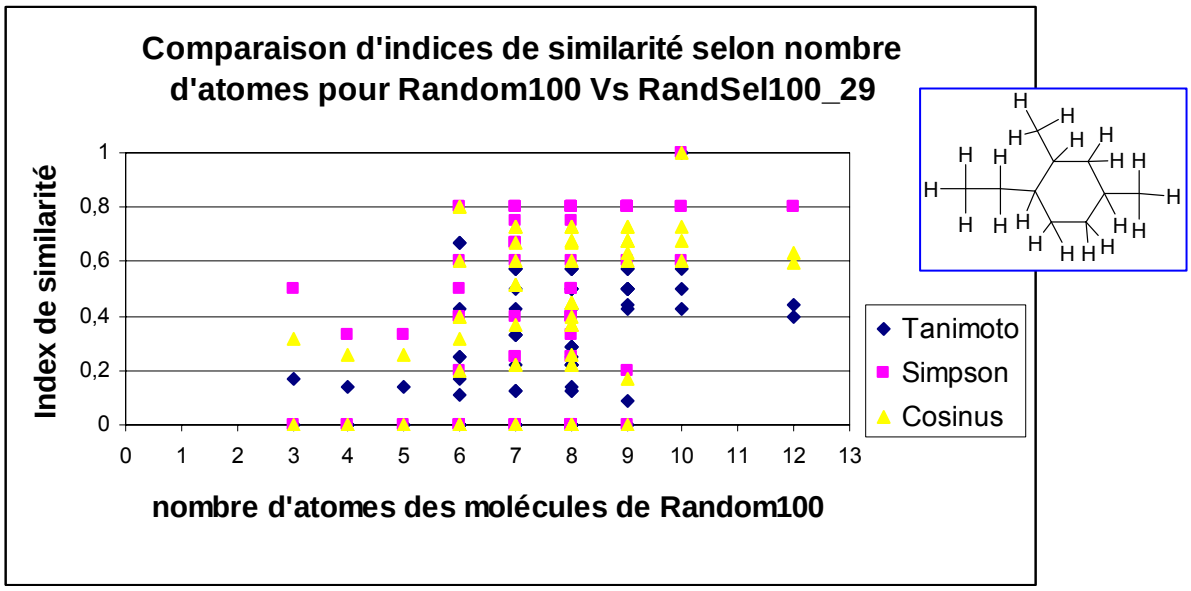
On remarquera que l'usage de la base « Random » pour faire des graphes de comparaison d'indices par rapport au nombre d'atomes génère des résultats assez différents, par rapport à la base « Zinc », particulièrement pour la distribution des données. Le rang du nombre d'atomes des molécules appartenant à la base « Random » est assez restreint. On sait que cette base est composée de petites molécules, et on peut observer que la plupart sont composées de 3 à 12 atomes.

Les valeurs de similarité se superposent particulièrement autour des molécules avec 6 et 8 atomes (RanSel100_16, RanSel100_29 et RanSel100_51) à l'exception de RandSel100_74, pour laquelle les valeurs de similarité sont en majorité nulles, dû à la rareté des fragments composant cette molécule. La distribution de valeurs élevées de similarité (Is entre 0,6 et 1) semble obéir à une règle définie : pour RandSel100_16 et RandSel100_29, la tendance est vers les molécules entre 6 et 9 atomes, tandis que pour RanSel100_51, c'est autour de 7 atomes. Les valeurs élevées de Is sont donc directement dépendantes de la taille de la molécule cible.

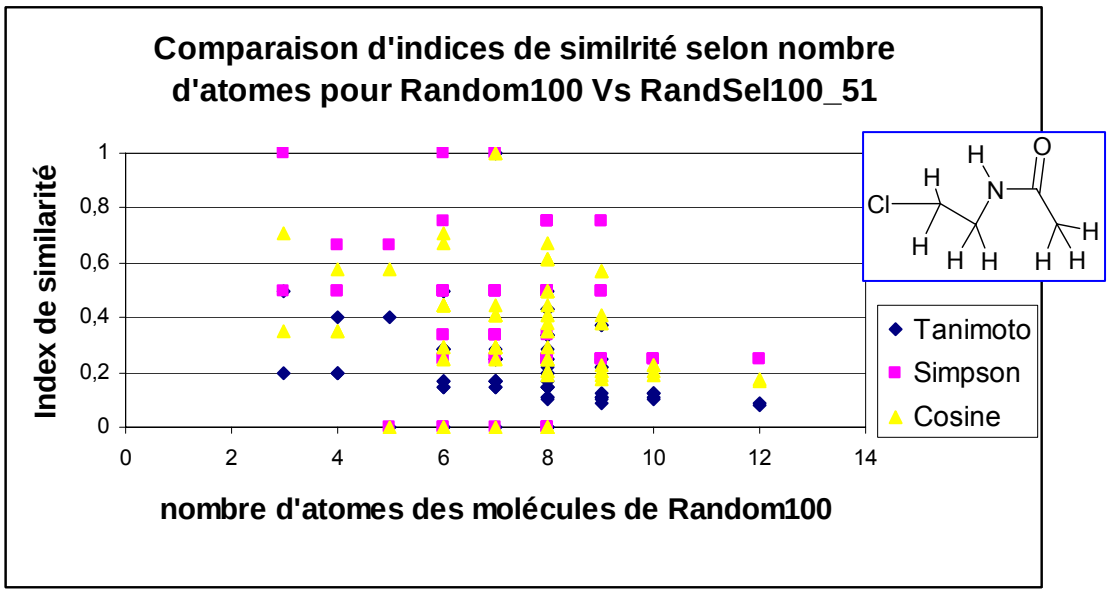
Graphique V.21 Comparaison de 3 mesures de similarité en fonction de la taille de molécules pour la base « Random » et la cible : RandSel100_16.mol



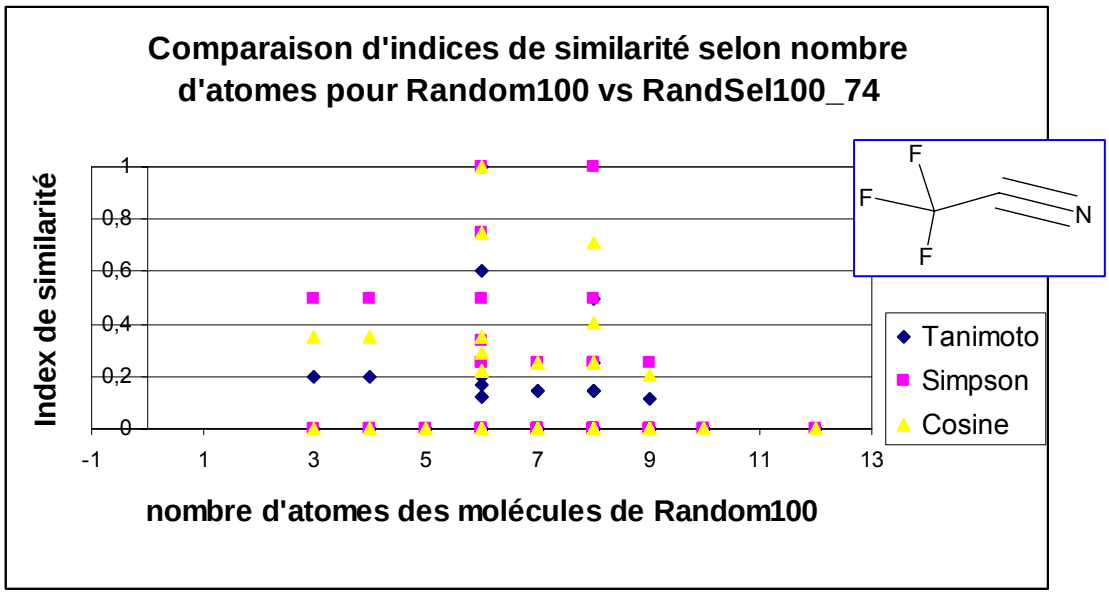
Graphique V.22 Comparaison de 3 mesures de similarité en fonction de la taille de molécules pour la base « Random » et la cible : RandSel100_29.mol



Graphique V.23 Comparaison de 3 mesures de similarité en fonction de la taille de molécules pour la base « Random » et la cible : RandSel100_51.mol



Graphique V.24 Comparaison de 3 mesures de similarité en fonction de la taille de molécules pour la base « Random » et la cible : RandSel100_74.mol



V.2 Analyse de type N-N

L'analyse de type N-N permet de calculer la mesure de similarité entre toutes les molécules appartenant à une base déterminée. Si on connaît les différences de tous les éléments d'une base, il est possible d'effectuer une analyse de l'espace de diversité de la base moléculaire.

V.2.1 Résultats avec la base « Zinc »

Lors du calcul structurel simple de type N-N avec la base « Zinc », le système construit une matrice carrée et diagonale de mesures de similarité. Les données correspondent aux 34 molécules appartenant à la base « Zinc ». Au total, il y a 1156 mesures de similarité correspondant à 34^2 combinaisons.

L'ensemble des molécules présente des valeurs de similarité assez hétérogènes. Pour l'index de Tanimoto (Tableau V.11), seules 3,63% de molécules présentent des indices de similarité supérieurs ou égaux à 0,8. Ceci n'est qu'une preuve de la diversité de la base. L'index de Simpson pour sa

part, compte 46,62% de molécules avec un index de similarité supérieur ou égal à 0,5, la plupart d'entre elles étant des isomorphismes non restrictifs sur l'enchaînement et la position des sous-structures trouvées. Les graphiques présentés dans le chapitre V.3 nous permettront d'évaluer d'une manière plus globale la distribution des molécules dans l'espace chimique.

Index	Mesures de Sim ≥ 0,5		Mesures de Sim ≥ 0,8		Mesures de Sim < 0,5	
Tanimoto	80	6,92%	42	3,63%	961	83,13%
Cosinus	232	20,07%	50	4,33%	810	70,07%
Simpson	539	46,62%	133	11,51%	503	43,51%

Tableau V.11 Nombre et % de molécules pour un analyse N-N sur la base « Zinc ».

Les résultats d'une partie des matrices de mesures de similarité pour les trois indices étudiés (Tanimoto, Simpson et Cosinus) sont montrés dans les tableaux V.12-V.14. Leurs graphiques 3D respectifs sont représentés par la suite.

On remarquera que tous les graphes sont symétriques par rapport au plan X-Y. Ceci est due à l'origine des données : une matrice carrée et diagonale issue de la comparaison multiple de toutes les molécules d'une base. La distribution des données a été faite de manière homogène pour éviter la superposition de points. Ceci peut être constaté dans le graphique V.25b.

Le nuage de mesures de similarité correspondant à l'index de Simpson est celui qui présente une distribution la plus homogène dans l'espace, par rapport aux deux autres indices. Mais pour qu'une base soit la plus diverse possible, la plupart des points devraient être dans le plan de la base. Ceci serait une garantie que la plupart des molécules ne se ressemblent pas les unes aux autres. Plus le nuage de points est vers le haut (Is~1), moins la base données sera hétérogène, le graphique V.26 étant un bon contre exemple.

D'autre part, on observe des paliers des mesures de similarité autour de 0.2, 0.4 et 0.6 en utilisant l'index de Simpson. Ceux ci correspondent à l'aire du maximum de correspondances des molécules de la base. Ce comportement, déjà observé lors des analyses 1-N, met en évidence la pauvre capacité de sélection de l'index de Simpson, particulièrement accentué lorsque l'on analyse des molécules de petite taille (voir graphique V.31).

Les différences entre les représentations graphiques 3D des indices, pour une même base de molécules, nous donnent l'information concernant la façon dont a été mené le calcul, mais aussi sur la composition de la base.

Tableau V.12 Analyse de Similarité/Diversité N-N d'une partie de la base « Zinc » avec Tanimoto.

	zinc_10	zinc_1146	zinc_1152	zinc_1326	zinc_1385	zinc_1514	zinc_1531	zinc_189	zinc_28	zinc_32	zinc_57	zinc_58	zinc_660	zinc_7
zinc_10	1,00	0,50	0,14	0,00	0,00	0,11	0,11	0,27	0,11	0,16	0,23	0,25	0,10	0,13
zinc_1146	0,50	1,00	0,16	0,12	0,14	0,12	0,12	0,30	0,12	0,50	0,25	0,27	0,11	0,14
zinc_1152	0,14	0,16	1,00	0,14	0,16	0,60	0,33	0,20	0,33	0,25	0,16	0,18	0,50	0,15
zinc_1326	0,00	0,12	0,14	1,00	0,80	0,25	0,66	0,07	0,11	0,16	0,06	0,07	0,22	0,06
zinc_1385	0,00	0,14	0,16	0,80	1,00	0,28	0,80	0,08	0,12	0,20	0,07	0,07	0,25	0,06
zinc_1514	0,11	0,12	0,60	0,25	0,28	1,00	0,42	0,16	0,25	0,16	0,14	0,15	0,57	0,13
zinc_1531	0,11	0,12	0,33	0,66	0,80	0,42	1,00	0,16	0,25	0,16	0,14	0,15	0,37	0,13
zinc_189	0,27	0,30	0,20	0,07	0,08	0,16	0,16	1,00	0,27	0,22	0,42	0,46	0,36	0,40
zinc_28	0,11	0,12	0,33	0,11	0,12	0,25	0,25	0,27	1,00	0,16	0,23	0,25	0,37	0,30
zinc_32	0,16	0,50	0,25	0,16	0,20	0,16	0,16	0,22	0,16	1,00	0,18	0,20	0,14	0,07
zinc_57	0,23	0,25	0,16	0,06	0,07	0,14	0,14	0,42	0,23	0,18	1,00	0,90	0,21	0,43
zinc_58	0,25	0,27	0,18	0,07	0,07	0,15	0,15	0,46	0,25	0,20	0,90	1,00	0,23	0,37
zinc_660	0,10	0,11	0,50	0,22	0,25	0,57	0,37	0,36	0,37	0,14	0,21	0,23	1,00	0,28
zinc_7	0,13	0,14	0,15	0,06	0,06	0,13	0,13	0,40	0,30	0,07	0,43	0,37	0,28	1,00

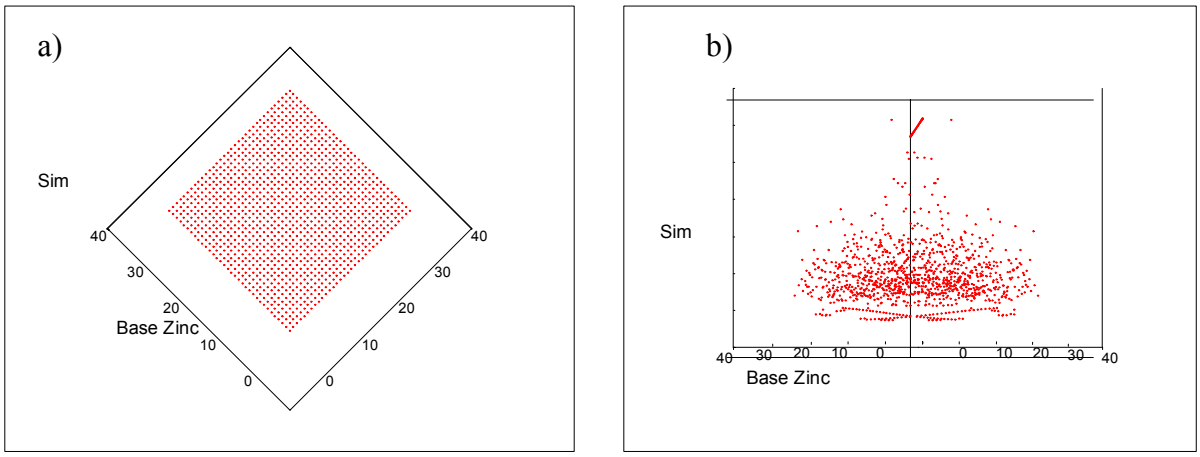
Tableau V.13 Analyse de Similarité/Diversité N-N d'une partie de la base « Zinc » avec Simpson.

	zinc_10	zinc_1146	zinc_1152	zinc_1326	zinc_1385	zinc_1514	zinc_1531	zinc_189	zinc_28	zinc_32	zinc_57	zinc_58	zinc_660	zinc_7
zinc_10	1,00	0,75	0,33	0,00	0,00	0,20	0,20	0,60	0,20	0,50	0,60	0,60	0,20	0,40
zinc_1146	0,75	1,00	0,33	0,25	0,25	0,25	0,25	0,75	0,25	1,00	0,75	0,75	0,25	0,50
zinc_1152	0,33	0,33	1,00	0,33	0,33	1,00	0,67	0,67	0,67	0,50	0,67	0,67	1,00	0,67
zinc_1326	0,00	0,25	0,33	1,00	1,00	0,40	0,80	0,20	0,20	0,50	0,20	0,20	0,40	0,20
zinc_1385	0,00	0,25	0,33	1,00	1,00	0,50	1,00	0,25	0,25	0,50	0,25	0,25	0,50	0,25
zinc_1514	0,20	0,25	1,00	0,40	0,50	1,00	0,60	0,40	0,40	0,50	0,40	0,40	0,80	0,40
zinc_1531	0,20	0,25	0,67	0,80	1,00	0,60	1,00	0,40	0,40	0,50	0,40	0,40	0,60	0,40
zinc_189	0,60	0,75	0,67	0,20	0,25	0,40	0,40	1,00	0,60	1,00	0,67	0,67	0,67	0,67
zinc_28	0,20	0,25	0,67	0,20	0,25	0,40	0,40	0,60	1,00	0,50	0,60	0,60	0,60	0,80
zinc_32	0,50	1,00	0,50	0,50	0,50	0,50	0,50	1,00	0,50	1,00	1,00	1,00	0,50	0,50
zinc_57	0,60	0,75	0,67	0,20	0,25	0,40	0,40	0,67	0,60	1,00	1,00	1,00	0,50	0,64
zinc_58	0,60	0,75	0,67	0,20	0,25	0,40	0,40	0,67	0,60	1,00	1,00	1,00	0,50	0,60
zinc_660	0,20	0,25	1,00	0,40	0,50	0,80	0,60	0,67	0,60	0,50	0,50	0,50	1,00	0,67
zinc_7	0,40	0,50	0,67	0,20	0,25	0,40	0,40	0,67	0,80	0,50	0,64	0,60	0,67	1,00

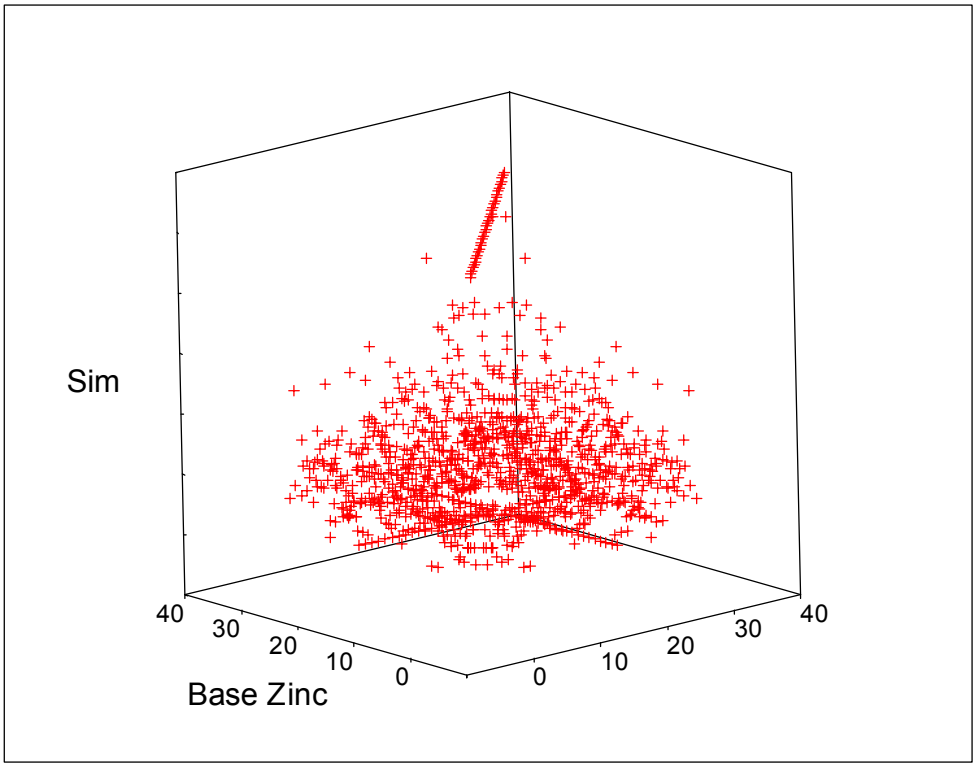
Tableau V.14 Analyse de Similarité/Diversité N-N d'une partie de la base « Zinc » avec Cosinus.

	zinc_10	zinc_1146	zinc_1152	zinc_1326	zinc_1385	zinc_1514	zinc_1531	zinc_189	zinc_28	zinc_32	zinc_57	zinc_58	zinc_660	zinc_7
zinc_10	1,00	0,67	0,26	0,00	0,00	0,20	0,20	0,45	0,20	0,32	0,40	0,42	0,18	0,26
zinc_1146	0,67	1,00	0,29	0,22	0,25	0,22	0,22	0,50	0,22	0,71	0,45	0,47	0,20	0,29
zinc_1152	0,26	0,29	1,00	0,26	0,29	0,77	0,52	0,38	0,52	0,41	0,35	0,37	0,71	0,33
zinc_1326	0,00	0,22	0,26	1,00	0,89	0,40	0,80	0,15	0,20	0,32	0,13	0,14	0,37	0,13
zinc_1385	0,00	0,25	0,29	0,89	1,00	0,45	0,89	0,17	0,22	0,35	0,15	0,16	0,41	0,14
zinc_1514	0,20	0,22	0,77	0,40	0,45	1,00	0,60	0,30	0,40	0,32	0,27	0,28	0,73	0,26
zinc_1531	0,20	0,22	0,52	0,80	0,89	0,60	1,00	0,30	0,40	0,32	0,27	0,28	0,55	0,26
zinc_189	0,45	0,50	0,38	0,15	0,17	0,30	0,30	1,00	0,45	0,47	0,60	0,63	0,54	0,58
zinc_28	0,20	0,22	0,52	0,20	0,22	0,40	0,40	0,45	1,00	0,32	0,40	0,42	0,55	0,52
zinc_32	0,32	0,71	0,41	0,32	0,35	0,32	0,32	0,47	0,32	1,00	0,43	0,45	0,29	0,20
zinc_57	0,40	0,45	0,35	0,13	0,15	0,27	0,27	0,60	0,40	0,43	1,00	0,95	0,37	0,61
zinc_58	0,42	0,47	0,37	0,14	0,16	0,28	0,28	0,63	0,42	0,45	0,95	1,00	0,39	0,55
zinc_660	0,18	0,20	0,71	0,37	0,41	0,73	0,55	0,54	0,55	0,29	0,37	0,39	1,00	0,47
zinc_7	0,26	0,29	0,33	0,13	0,14	0,26	0,26	0,58	0,52	0,20	0,61	0,55	0,47	1,00

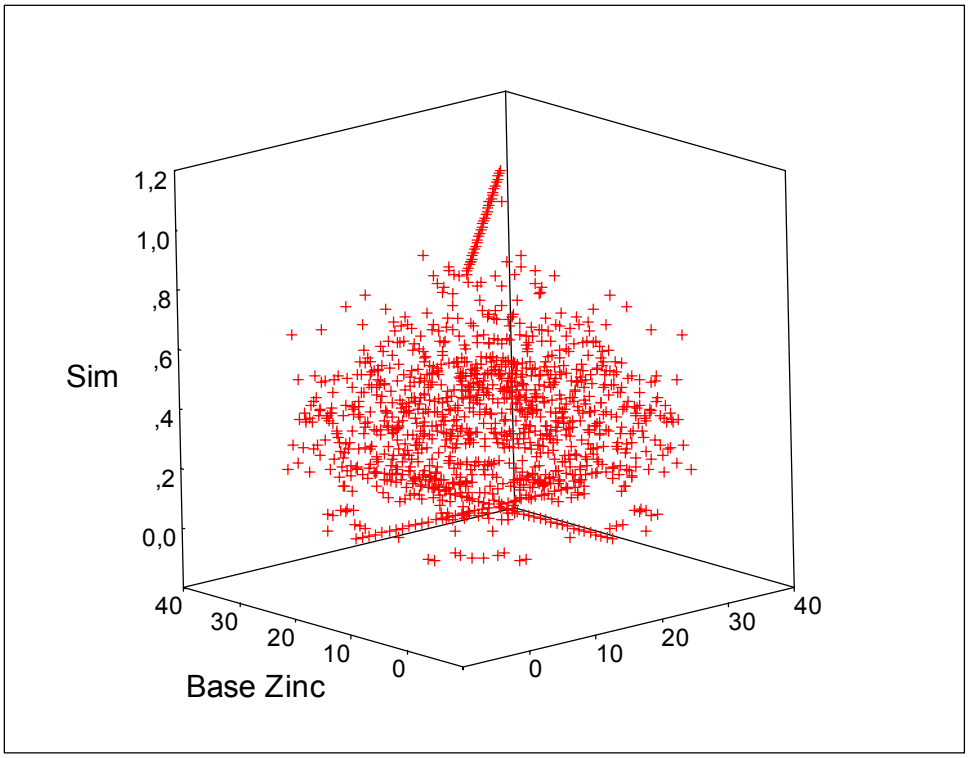
Graphique V.25 Différents vues des graphiques N-N : a) Vue « de haut » du graphique 3D. La distribution des données est homogène. b) Vue « de face » du graphique 3D.



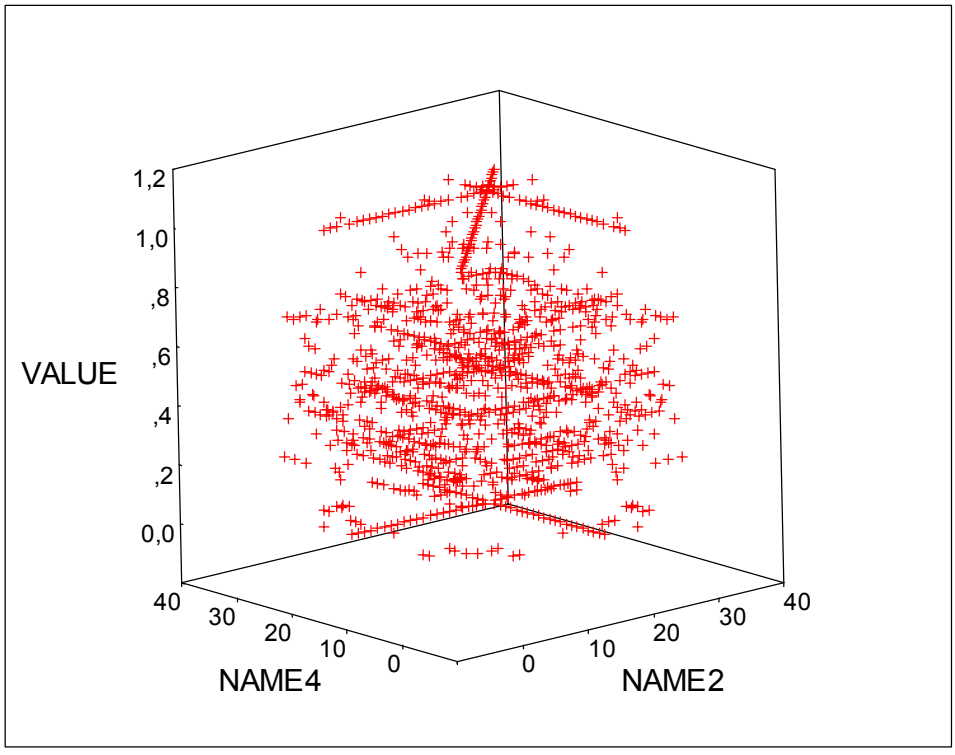
Graphique V.26 Analyse de Similarité/Diversité N-N avec la base « Zinc ». Indice de Tanimoto.



Graphique V.27 Analyse de Similarité/Diversité N-N avec la base « Zinc ». Indice de Cosinus.



Graphique V.28 Analyse de Similarité/Diversité N-N avec la base « Zinc ». Indice de Simpson.



V.2.2 Résultats avec la base « Random »

Pour l'analyse N-N avec les molécules de la base « Random », nous avons obtenu au total 5929 mesures de similarité correspondant à 77^2 combinaisons (car Random est composé de 77 molécules). Les indices de Tanimoto et de Cosinus donnent les meilleurs résultats en termes de sélectivité des requêtes. Toutefois les proportions restent très inégales (140 et 279 molécules respectivement pour une mesure de similarité ≥ 0.8, soit 2.36% et 4.71% de la base totale).

Index	Mesures de Sim ≥ 0,5		Mesures de Sim ≥ 0,8		Mesures de Sim < 0,5	
Tanimoto	820	13,83%	140	2,36%	3265	55,07%
Cosinus	1617	27,27%	279	4,71%	2468	41,63%
Simpson	2796	47,16%	676	11,40%	1289	21,74%

Tableau V.15 Nombre et % de molécules pour une analyse N-N sur la base « Zinc ».

A l'instar des résultats montrés dans la section précédente, les résultats d'une partie des matrices de similarité pour les trois indices étudiés (Tanimoto, Simpson et Cosinus) sont montrés dans les tableaux V.16-V.18.

On remarque que l'augmentation du nombre de mesures part rapport à l'analyse N-N avec la base « Zinc », améliore la visibilité de l'ensemble des données. Le nuage de mesures de similarité (tous indices confondus) présente une distribution assez hétérogène dans l'espace, par rapport à la base « Zinc ». Ceci se traduit par une pauvre diversité expliquée en partie dans l'origine des données composant la base « Random » : un ensemble de molécules appartenant à des groupes d'isomères de formule fixe, mais avec des constructions structurales variables.

La nature de la base « Random » (composée principalement par des groupes de petites molécules regroupées par taille et nombre d'atomes) pourrait également expliquer la présence des nombreux paliers observés dans le graphique V.31, même si ce comportement a déjà été présent lors de l'analyse des molécules de la base « Zinc ».

Tableau V.16 Analyse N-N d'une partie de la base « Random ». Indice de Tanimoto.

	RS100_1	RS100_10	RS100_100	RS100_11	RS100_13	RS100_14	RS100_15	RS100_16	RS100_17	RS100_18	RS100_19	RS100_2	RS100_20	RS100_21
RandSel100_1	1,00	0,00	0,33	0,00	0,50	0,00	0,60	0,80	0,00	0,75	0,00	0,60	1,00	0,25
RandSel100_10	0,00	1,00	0,33	0,25	0,50	0,50	0,20	0,00	0,50	0,25	0,50	0,17	0,00	0,25
RandSel100_100	0,33	0,33	1,00	0,33	0,00	0,00	0,00	0,67	0,33	0,33	0,50	0,33	0,50	0,33
RandSel100_11	0,00	0,25	0,33	1,00	0,00	0,00	0,00	0,00	0,25	0,00	0,50	0,25	0,00	0,25
RandSel100_13	0,50	0,50	0,00	0,00	1,00	0,50	1,00	0,50	0,50	1,00	0,00	0,50	0,00	0,50
RandSel100_14	0,00	0,50	0,00	0,00	0,50	1,00	0,50	0,00	0,50	0,50	0,00	0,00	0,00	0,00
RandSel100_15	0,60	0,20	0,00	0,00	1,00	0,50	1,00	0,80	0,25	0,75	0,00	0,80	0,00	0,25
RandSel100_16	0,80	0,00	0,67	0,00	0,50	0,00	0,80	1,00	0,00	0,75	0,00	0,67	0,50	0,25
RandSel100_17	0,00	0,50	0,33	0,25	0,50	0,50	0,25	0,00	1,00	0,25	0,50	0,25	0,00	0,50
RandSel100_18	0,75	0,25	0,33	0,00	1,00	0,50	0,75	0,75	0,25	1,00	0,00	0,50	0,50	0,25
RandSel100_19	0,00	0,50	0,50	0,50	0,00	0,00	0,00	0,00	0,50	0,00	1,00	0,50	0,00	0,50
RandSel100_2	0,60	0,17	0,33	0,25	0,50	0,00	0,80	0,67	0,25	0,50	0,50	1,00	0,00	0,50
RandSel100_20	1,00	0,00	0,50	0,00	0,00	0,00	0,00	0,50	0,00	0,50	0,00	0,00	1,00	0,00
RandSel100_21	0,25	0,25	0,33	0,25	0,50	0,00	0,25	0,25	0,50	0,25	0,50	0,50	0,00	1,00

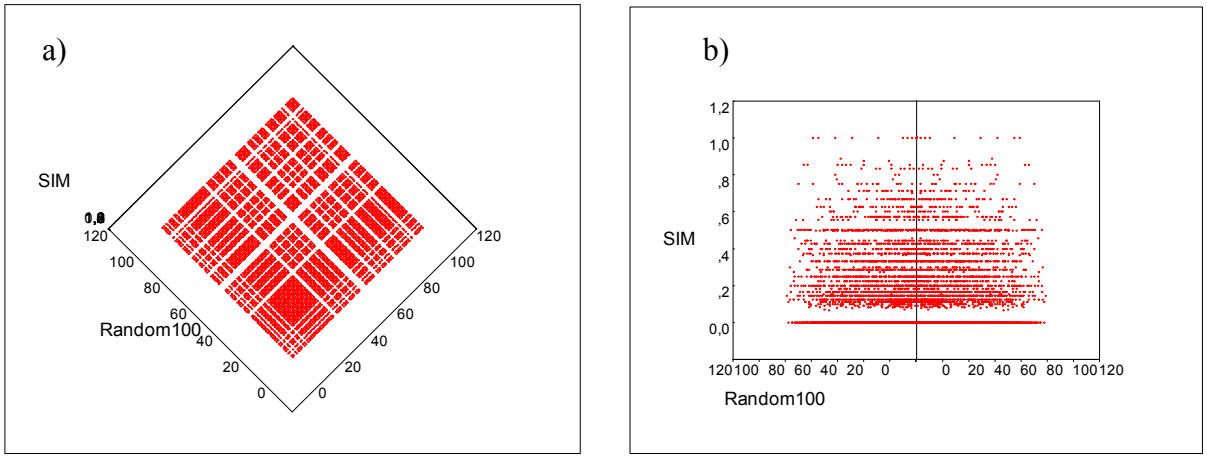
Tableau V.17 Analyse N-N d'une partie de la base « Random ». Indice de Cosinus.

	RS100_1	RS100_10	RS100_100	RS100_11	RS100_13	RS100_14	RS100_15	RS100_16	RS100_17	RS100_18	RS100_19	RS100_2	RS100_20	RS100_21
RandSel100_1	1,00	0,00	0,26	0,00	0,32	0,00	0,60	0,73	0,00	0,67	0,00	0,51	0,63	0,22
RandSel100_10	0,00	1,00	0,24	0,20	0,29	0,29	0,18	0,00	0,41	0,20	0,29	0,15	0,00	0,20
RandSel100_100	0,26	0,24	1,00	0,29	0,00	0,00	0,00	0,47	0,29	0,29	0,41	0,22	0,41	0,29
RandSel100_11	0,00	0,20	0,29	1,00	0,00	0,00	0,00	0,00	0,25	0,00	0,35	0,19	0,00	0,25
RandSel100_13	0,32	0,29	0,00	0,00	1,00	0,50	0,63	0,29	0,35	0,71	0,00	0,27	0,00	0,35
RandSel100_14	0,00	0,29	0,00	0,00	0,50	1,00	0,32	0,00	0,35	0,35	0,00	0,00	0,00	0,00
RandSel100_15	0,60	0,18	0,00	0,00	0,63	0,32	1,00	0,73	0,22	0,67	0,00	0,68	0,00	0,22
RandSel100_16	0,73	0,00	0,47	0,00	0,29	0,00	0,73	1,00	0,00	0,61	0,00	0,62	0,29	0,20
RandSel100_17	0,00	0,41	0,29	0,25	0,35	0,35	0,22	0,00	1,00	0,25	0,35	0,19	0,00	0,50
RandSel100_18	0,67	0,20	0,29	0,00	0,71	0,35	0,67	0,61	0,25	1,00	0,00	0,38	0,35	0,25
RandSel100_19	0,00	0,29	0,41	0,35	0,00	0,00	0,00	0,00	0,35	0,00	1,00	0,27	0,00	0,35
RandSel100_2	0,51	0,15	0,22	0,19	0,27	0,00	0,68	0,62	0,19	0,38	0,27	1,00	0,00	0,38
RandSel100_20	0,63	0,00	0,41	0,00	0,00	0,00	0,00	0,29	0,00	0,35	0,00	0,00	1,00	0,00
RandSel100_21	0,22	0,20	0,29	0,25	0,35	0,00	0,22	0,20	0,50	0,25	0,35	0,38	0,00	1,00

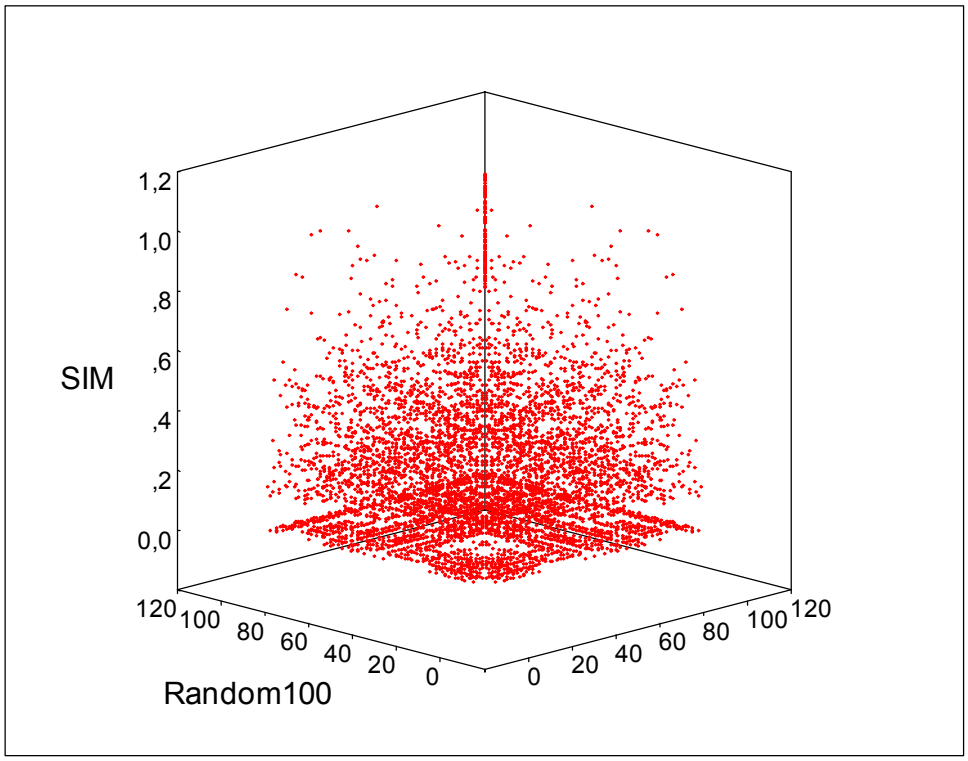
Tableau V.18 Analyse N-N d'une partie de la base « Random ». Indice de Simpson.

	RS100_1	RS100_10	RS100_100	RS100_11	RS100_13	RS100_14	RS100_15	RS100_16	RS100_17	RS100_18	RS100_19	RS100_2	RS100_20	RS100_21
RandSel100_1	1,00	0,00	0,33	0,00	0,50	0,00	0,60	0,80	0,00	0,75	0,00	0,60	1,00	0,25
RandSel100_10	0,00	1,00	0,33	0,25	0,50	0,50	0,20	0,00	0,50	0,25	0,50	0,17	0,00	0,25
RandSel100_100	0,33	0,33	1,00	0,33	0,00	0,00	0,00	0,67	0,33	0,33	0,50	0,33	0,50	0,33
RandSel100_11	0,00	0,25	0,33	1,00	0,00	0,00	0,00	0,00	0,25	0,00	0,50	0,25	0,00	0,25
RandSel100_13	0,50	0,50	0,00	0,00	1,00	0,50	1,00	0,50	0,50	1,00	0,00	0,50	0,00	0,50
RandSel100_14	0,00	0,50	0,00	0,00	0,50	1,00	0,50	0,00	0,50	0,50	0,00	0,00	0,00	0,00
RandSel100_15	0,60	0,20	0,00	0,00	1,00	0,50	1,00	0,80	0,25	0,75	0,00	0,80	0,00	0,25
RandSel100_16	0,80	0,00	0,67	0,00	0,50	0,00	0,80	1,00	0,00	0,75	0,00	0,67	0,50	0,25
RandSel100_17	0,00	0,50	0,33	0,25	0,50	0,50	0,25	0,00	1,00	0,25	0,50	0,25	0,00	0,50
RandSel100_18	0,75	0,25	0,33	0,00	1,00	0,50	0,75	0,75	0,25	1,00	0,00	0,50	0,50	0,25
RandSel100_19	0,00	0,50	0,50	0,50	0,00	0,00	0,00	0,00	0,50	0,00	1,00	0,50	0,00	0,50
RandSel100_2	0,60	0,17	0,33	0,25	0,50	0,00	0,80	0,67	0,25	0,50	0,50	1,00	0,00	0,50
RandSel100_20	1,00	0,00	0,50	0,00	0,00	0,00	0,00	0,50	0,00	0,50	0,00	0,00	1,00	0,00
RandSel100_21	0,25	0,25	0,33	0,25	0,50	0,00	0,25	0,25	0,50	0,25	0,50	0,50	0,00	1,00

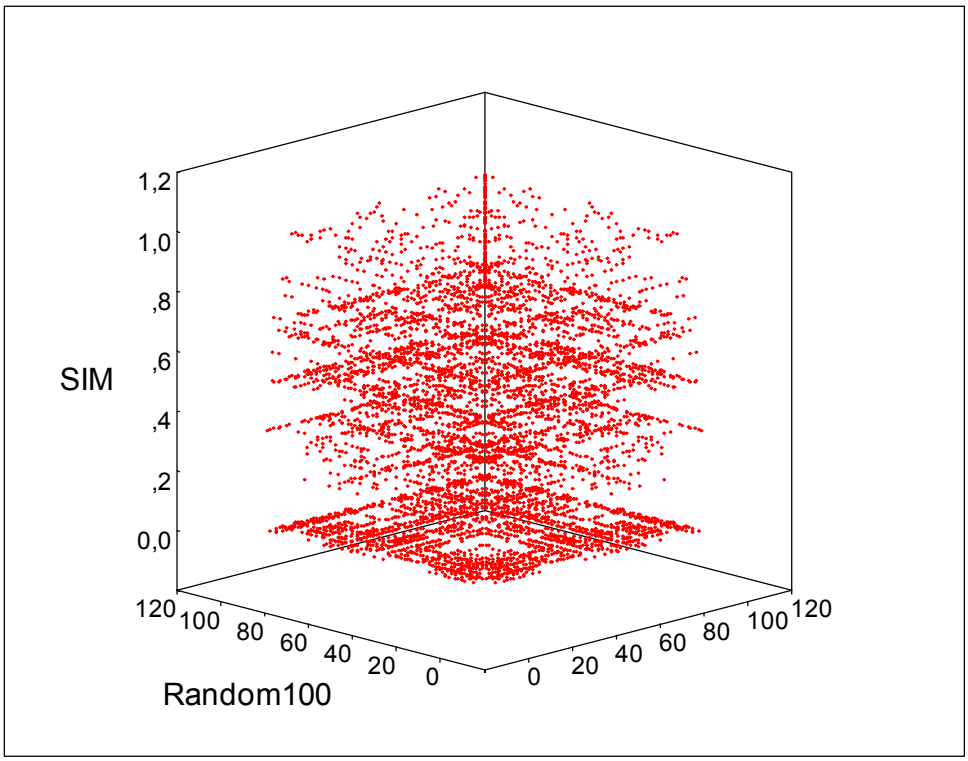
Graphique V.29 Deux vues des graphiques N-N : a) Vue « de haut » du graphique 3D. La distribution des données est homogène. b) Vue « de face » du graphique 3D. La symétrie du dessin est due à l'origine des données (matrice diagonale, donc symétrique). Le nombre de points est bien supérieur à ceux du graphique V.25.



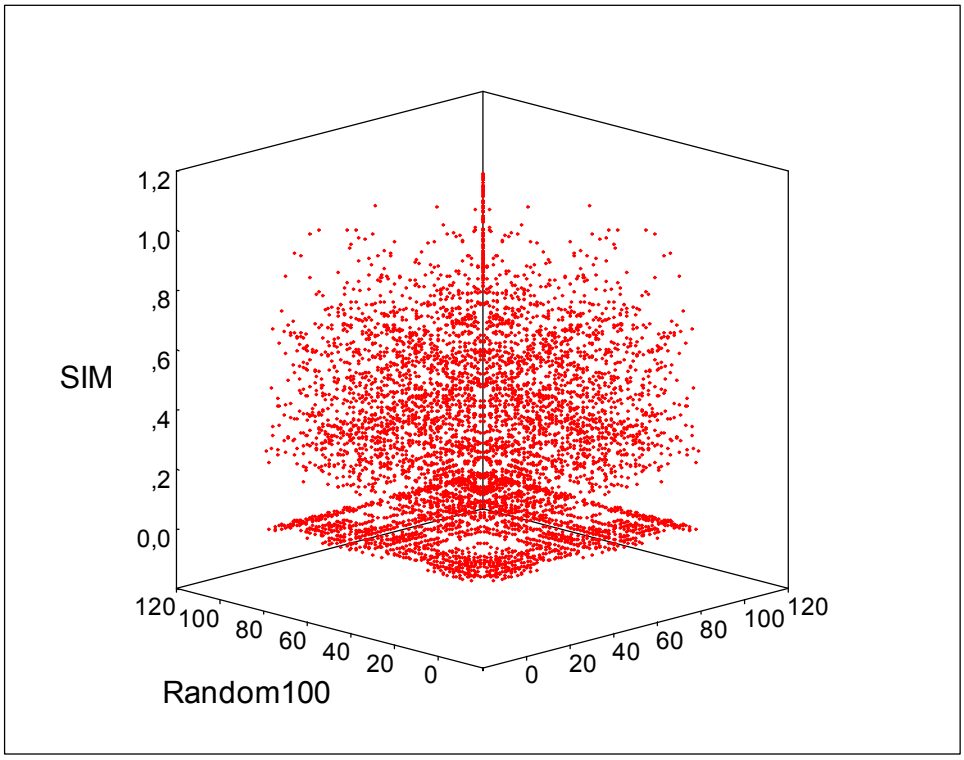
Graphique V.30 Analyse de Similarité/Diversité N-N avec « Random ». Indice de Tanimoto.



Graphique V.31 Analyse de Similarité/Diversité N-N avec « Random ». Indice de Simpson.



Graphique V.32 Analyse de Similarité/Diversité N-N avec « Random ». Indice de Cosinus.

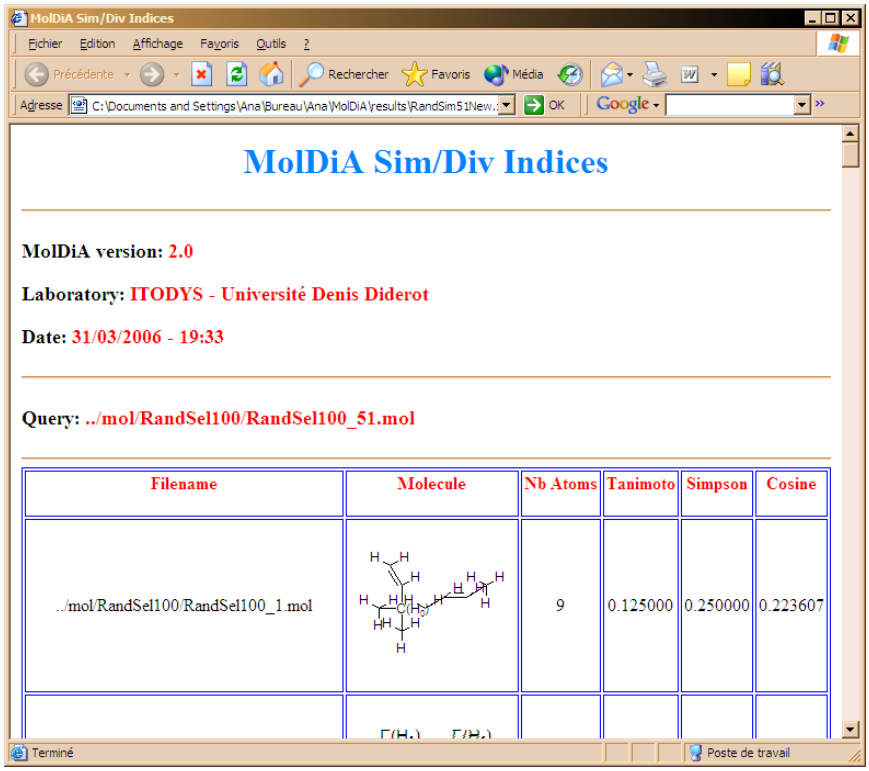


V.2.3 Aperçu des résultats structurés et présentés avec XML

Comme nous l'avons indiqué auparavant, les vecteurs de sous-structures pour chaque fragment, ainsi que les fichiers résultats pour les indices choisis, sont structurés en XML et peuvent également être présentés en utilisant une feuille de style XSL. Cette présentation automatise l'élaboration des tableaux de résultats et inclus une présentation graphique de la molécule (en 2D ou 3D), pour faciliter l'analyse des résultats obtenus. Des copies d'écran des fichiers XML ouverts avec Internet Explorer en utilisant le fichier XSL fourni dans l'outil sont données par la suite. Veuillez noter que tous ces fichiers formatés sont générés automatiquement et peuvent être ouverts sous Excel, ou tout outil graphique qui supporte les fichiers XML (la plupart aujourd'hui). Des fragments du code original sont également montrés pour mettre en évidence l'usage des balises dans la structuration de l'information qui sera ensuite utilisée pour effectuer la présentation des données.

Figure V.3 Copies d'écran de fichiers de mesures de similarité 1-N ouverts avec Internet Explorer.

A droite un extrait du fichier XML de similarité/diversité original.




```
<?xml version='1.0' encoding='iso-
8859-1' standalone='no'?>
<?xml-stylesheet type='text/xsl'
href='MolDiA.xsl'?>
<MolDiA version='2.0'
laboratory='ITODYS - Université Denis
Diderot' date='31/03/2006 - 19:33'>
<Query fileName='RandSel100_51.mol'>
<Results>
<Test fileName='RandSel100_1.mol'>
<Molecule atom='9'/>
<Index Tanimoto='0.125000'
Simpson='0.250000'
Cosinus='0.223607'/>
</Test>
<Test fileName='RandSel100_10.mol'>
<Molecule atom='9'/>
<Index Tanimoto='0.250000'
Simpson='0.500000'
Cosinus='0.408248'/>
</Test>
...
</Results>
</Query>
</MolDiA>
```

Figure V.4 Copies d'écran de fichiers de mesures de similarité N-N ouverts avec Internet Explorer.

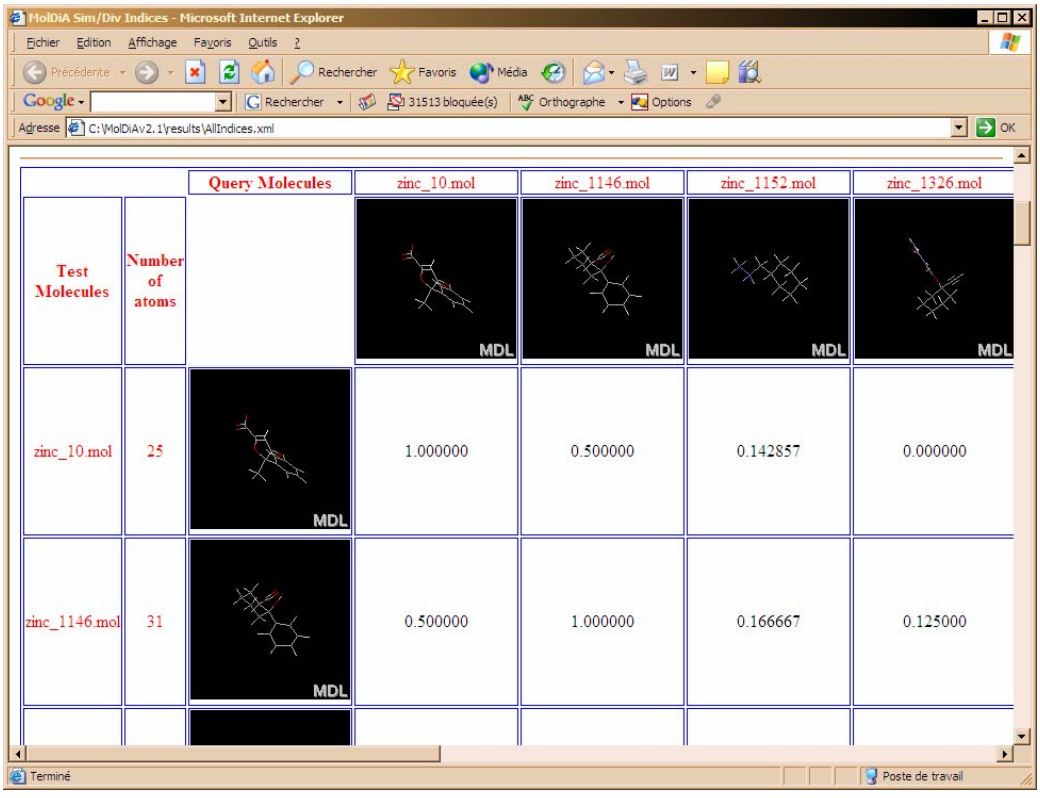
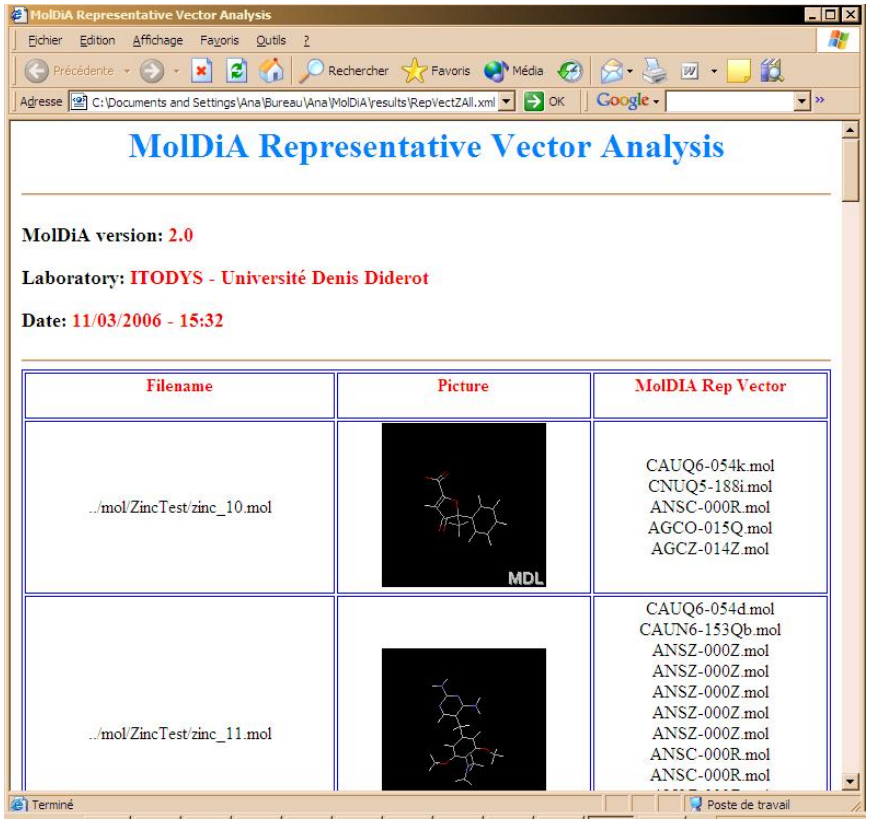


Figure V.5 Copies d'écran de fichiers de vecteurs résultats ouverts avec Internet Explorer.

A droite un extrait du fichier XML de vecteur résultat original.




```
<?xml version='1.0' encoding='iso-
8859-1' standalone='no'?>
<?xml-stylesheet type='text/xsl'
href='RepVec.xsl'?>
<MolDiA version='2.0'
laboratory='ITODYS - Université Denis
Diderot' date='11/03/2006 - 15:32'>
<Molecule fileName='zinc_10.mol'>
<ExpRepVector>
<Frag ref='CAUQ6-054k.mol'/>
<Frag ref='CNUQ5-188i.mol'/>
<Frag ref='ANSC-000R.mol'/>
<Frag ref='AGCO-015Q.mol'/>
<Frag ref='AGCZ-014Z.mol'/>
</ExpRepVector>
</Molecule>
<Molecule fileName='zinc_11.mol'>
<ExpRepVector>
<Frag ref='CAUQ6-054d.mol'/>
<Frag ref='CAUN6-153Qb.mol'/>
<Frag ref='ANSZ-000Z.mol'/>
<Frag ref='ANSZ-000Z.mol'/>
<Frag ref='ANSZ-000Z.mol'/>
<Frag ref='ANSZ-000Z.mol'/>
<Frag ref='ANSZ-000Z.mol'/>
<Frag ref='ANSC-000R.mol'/>
<Frag ref='ANSC-000R.mol'/>
<Frag ref='ANSZ-000Z.mol'/>
<Frag ref='ANSZ-000Z.mol'/>
<Frag ref='ANSZ-000Z.mol'/>
<Frag ref='ANSZ-000Z.mol'/>
</ExpRepVector>
</Molecule>
...
</MolDiA>
```

V.3 Evaluation de l'outil

Plusieurs méthodes ont été implémentées pour tester la capacité de l'outil à décrire de manière correcte l'information chimique de nature structurale contenue dans les molécules. Ces méthodes, tirées des mathématiques statistiques, nous permettront de mieux juger dans l'ensemble, les résultats des vecteurs descripteurs fournis par l'outil. De la même manière, des résultats erronés ont été analysés de manière à pouvoir tirer des conclusions sur le rang d'utilisation de l'outil ainsi que sur les forces et les faiblesses du programme.

V.3.1 Précision, rappel, et *F-mesure*, pour la base « Zinc »

Nous proposons une méthode qui permet d'évaluer l'erreur et l'exactitude d'un vecteur représentatif déterminé, calculé par notre outil. Seront donc comparées, une structure vectorielle générée par l'outil qu'on appellera « Sg » (ou vecteur «expérimental»), et la structure vectorielle théorique qu'on appellera « St » (ou vecteur «théorique»), puisque construite manuellement en observant minutieusement la structure de la molécule. Trois formules de comparaison sont implémentées : la précision, le rappel, et la F-mesure. Leurs définition et utilisation dans le cadre d'évaluation de l'outil ont déjà été l'objet du chapitre IV.2.2. Ici, nous montrerons simplement les équations pour leur calcul, ainsi que les résultats dans le tableau V.19

$$\text{Pr}\,ecision = \frac{St \cap Sg}{Sg} \tag{1}$$

$$Rappel = \frac{St \cap Sg}{St} \tag{2}$$

$$\text{F - measure} = \frac{2 \times \left(\text{Pr}\,ecision \times Rappel\right)}{\text{Pr}\,ecision + Rappel} \tag{3}$$

Où :

« St » est la structure vectorielle théorique construite manuellement pour évaluer l'outil.

« Sg » est la structure vectorielle générée par l'outil, donc expérimentale.

Tableau V.19 Précision, Rappel et F-Mesure entre les vecteurs représentatifs générés par l'outil pour les molécules de la base « Zinc » et les vecteurs théoriques attendus.

Filename	Precision	Recall	F-Measure
zinc_10.mol	0.600000	0.600000	0.600000
zinc_11.mol	0.846154	0.846154	0.846154
zinc_17.mol	1.000000	0.600000	0.750000
zinc_18.mol	1.000000	1.000000	1.000000
zinc_28.mol	1.000000	0.833333	0.909091
zinc_32.mol	1.000000	0.500000	0.666667
zinc_38.mol	1.000000	1.000000	1.000000
zinc_41.mol	0.600000	1.000000	0.750000
zinc_48.mol	1.000000	0.777778	0.875000
zinc_52.mol	0.714286	0.454545	0.555556
zinc_53.mol	0.600000	0.750000	0.666667
zinc_57.mol	1.000000	1.000000	1.000000
zinc_58.mol	1.000000	1.000000	1.000000
zinc_7.mol	1.000000	0.923077	0.960000
zinc_1036.mol	1.000000	0.857143	0.923077
zinc_1037.mol	0.833333	1.000000	0.909091
zinc_1145.mol	1.000000	0.600000	0.750000
zinc_1146.mol	1.000000	0.600000	0.750000
zinc_1152.mol	1.000000	0.750000	0.857143
zinc_1326.mol	1.000000	0.833333	0.909091
zinc_1385.mol	1.000000	0.800000	0.888889
zinc_1514.mol	1.000000	0.833333	0.909091
zinc_1527.mol	0.833333	1.000000	0.909091
zinc_1531.mol	1.000000	1.000000	1.000000
zinc_189.mol	1.000000	1.000000	1.000000
zinc_249.mol	1.000000	1.000000	1.000000
zinc_252.mol	1.000000	0.750000	0.857143
zinc_370.mol	1.000000	1.000000	1.000000
zinc_447.mol	1.000000	1.000000	1.000000
zinc_530.mol	1.000000	0.200000	0.333333
zinc_660.mol	1.000000	1.000000	1.000000
zinc_707.mol	1.000000	0.833333	0.909091
zinc_888.mol	1.000000	0.600000	0.750000
zinc_948.mol	0.857143	0.857143	0.857143

Dans le tableau suivant, les résultats en % pour la Précision, le Rappel et la F-Mesure des vecteurs descripteurs générés pour les molécules de la base « Zinc » sont récapitulés. On observe dans ce tableau que, malgré une haute précision pour l'ensemble des données, le rappel n'atteint pas toujours 100%. Ceci est une conséquence de la génération des vecteurs expérimentaux avec plus ou moins de fragments que ceux attendus.

	Nombre de molécules ayant 100% de correspondance St- Sg		Nombre de molécules ayant plus de 80% de correspondance St-Sg		Nombre de molécules ayant moins de 50% de correspondance St-Sg	
Précision	26	76,47%	30	88,24%	0	0%
Rappel	10	29,41%	22	64,71%	3	8,82%
F-Measure	10	29,41%	24	70,59%	2	5,88%

Tableau V.20 Tableau récapitulatif des correspondances St-Sg en fonction du nombre de molécules.

Deux cas sont donc possibles :

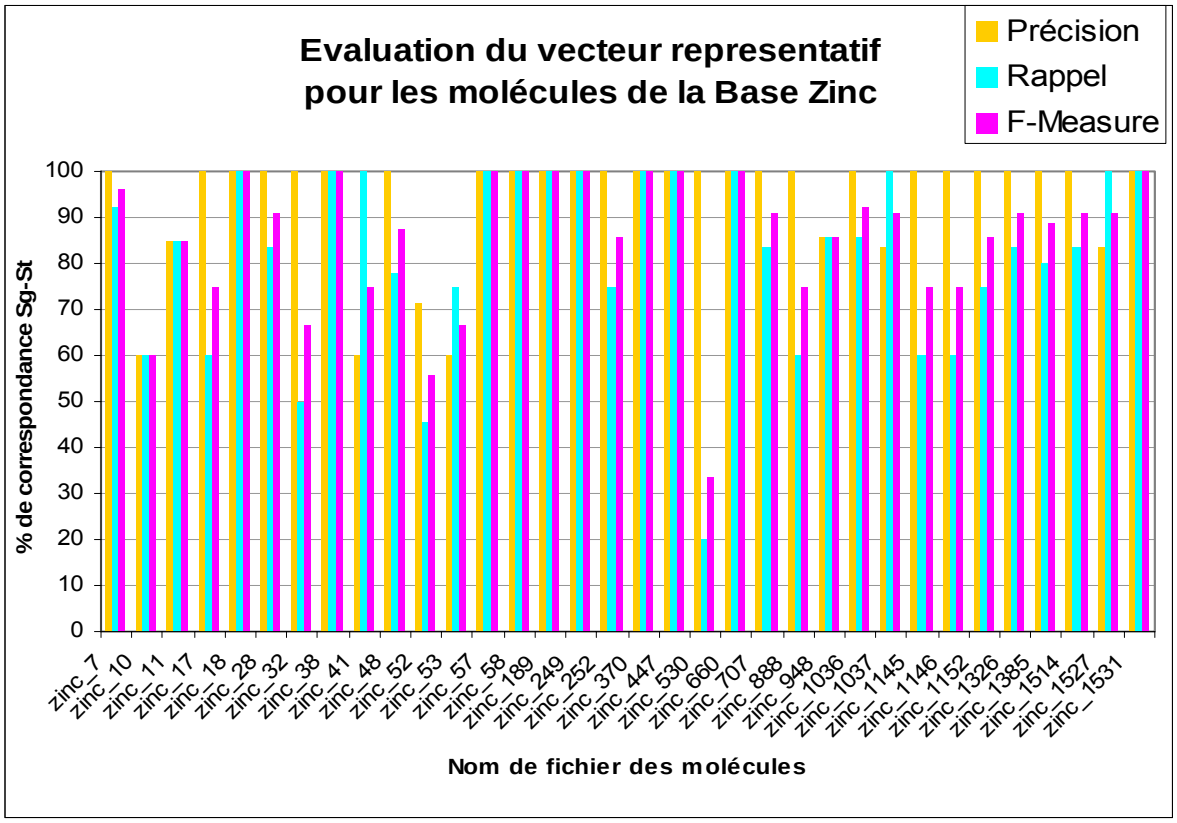
- Si le vecteur descripteur a plus de fragments, et si ceux-ci sont des doublons, l'effet sur la comparaison est de donner plus d'importance à cette structure, car elle aura plus de chances d'être détectée que les autres. L'implémentation de filtres devrait éliminer ce problème. Si ceux-ci ne sont pas des doublons, ils sont alors des fragments approximatifs à la sous-structure voulue. Ceci est dû à l'usage d'atomes génériques. L'effet sur la comparaison est de diminuer les correspondances totales avec la cible. Pour corriger ce problème, l'implémentation d'un niveau de comparaison « flou » (valable seulement si la comparaison exacte par nom de fichier échoue) qui prend en compte seulement la classe à laquelle appartient la sous-structure (cyclohexane, carbonyle, pyrimidine…) indépendamment de sa connectivité.

- Si le vecteur descripteur à moins de fragments, le plus probable est qu'il y a une ou plusieurs sous-

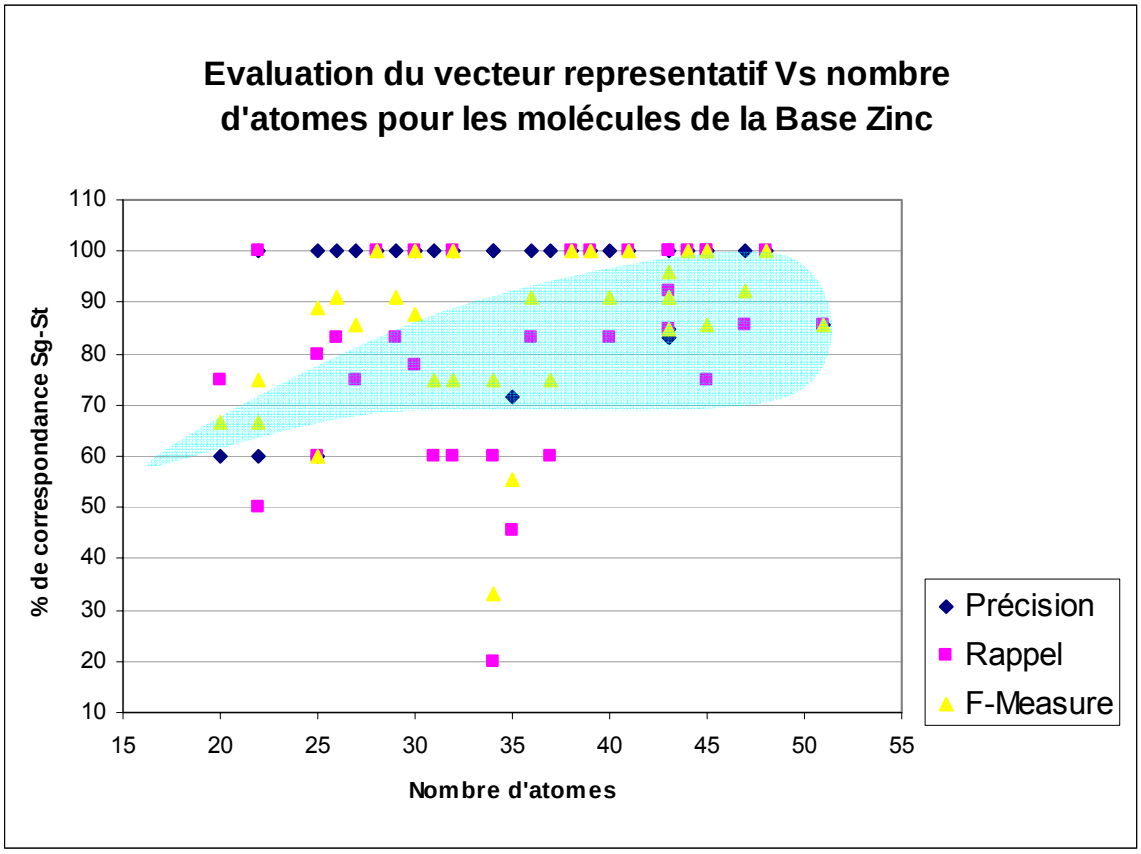
structures absentes de la base de fragments. Il suffit donc de modifier la base de fragment en rajoutant cette structure, mais aussi modifier l'index de fragments XML, le fichier de règles XML, et tout autre document susceptible de contenir l'information de la FragDB.

On remarque que le rapport entre la précision et le rappel pour 70,59% des molécules est de plus de 80%. Ces valeurs sont satisfaisantes pour les molécules de caractère complexe originaires de la base « Zinc » (voir Graphique V.33). D'une manière générale, la tendance est d'avoir de meilleurs descripteurs avec des plus grandes molécules (Graphique V.34), même si la tendance est plus marquée pour des molécules ayant moins de 10 atomes.

Graphique V.33 Comparaison de la précision, le rappel et la F-Mesure pour les molécules de « Zinc»



Graphique V.34 Comparaison de la précision, le rappel et la F-Mesure *vs* nombre d'atomes pour les molécules appartenant à la base « Zinc». La zone bleue montre la tendance observée des mesures de précision et de rappel avec l'augmentation du nombre d'atomes des molécules testées.



V.3.2 Etude des faux isomorphismes pour des mesures de similarité N-N

Dans notre cadre d'évaluation, un faux isomorphisme est défini comme une comparaison entre deux descripteurs qui malgré sa valeur unitaire, ne correspond pas à une correspondance exacte du point de vue moléculaire.

Ceci peut avoir plusieurs raisons :

- Les descripteurs structuraux définis pour la construction de l'outil de criblage virtuel ne prennent pas en compte la position de liaison de la sous-structure. Des isomères structuraux et optiques peuvent donc donner des valeurs de similarité unitaires.

- Si le vecteur descripteur a plus ou moins d'éléments qu'il devrait (voir section V.4.1 pour une liste non exhaustive de possibilités), des correspondances peuvent s'avérer erronées.

Les deux tableaux ci-dessous nous montrent que, indépendamment de la base utilisée, il y a moins de 1% de faux isomorphismes si on utilise les indices de Tanimoto et de Cosinus, et entre 2 et 3% avec le coefficient de Simpson (où plus de la moitié appartient à une même molécule). On remarquera également que pour la base « Random », presque la moitié des molécules impliquées dans des faux isomorphismes ont moins de 7 atomes, tout indices confondus, et que près de 100% avaient moins de 10 atomes. Ceci confirme la tendance déjà observée dans les analyses N - N des bases moléculaires.

Zinc Index	Mesures de similarité faussement identiques (faux Is = 1)		Molécules concernées appartenant à une même cible	
Tanimoto	1	0,09%	0	0%
Cosinus	1	0,09%	0	0%
Simpson	39	3,73%	28	73,68%

Tableau V.21 Tableau récapitulatif des faux isomorphismes observés dans la base « Zinc ».

Random Index	Mesures de similarité faussement identiques (faux Is = 1)		Molécules concernées ayant ≤ 7 atomes		Molécules concernées ayant ≤ 10 atomes	
Tanimoto	8	0,13%	4	50%	8	100%
Cosinus	8	0,13%	4	50%	8	100%
Simpson	113	1,90%	50	44,24%	111	98,23%

Tableau V.22 Tableau récapitulatif des faux isomorphismes observés dans la base « Random ».

V.3.3 Limites et avantages de l'outil

Le cadre d'application et les limites de l'approche peuvent être résumés en quelques points :

- La base de fragments est pour l'instant limitée à 700 sous-structures. Ceci est toutefois extensible.

- Le traitement des molécules est fait strictement sous format MDL .mol bien défini (voir annexe). Si le fichier .mol d'une molécule n'est pas bien défini (erreur dans la structure, absence d'une marque fin de fichier, etc.) ceci peut empêcher l'outil de traiter la molécule et interrompe le déroulement du calcul.

- Les informations moléculaires telles qu'elles sont définies dans les fichiers .mol sont strictement respectées lors du calcul (par exemple, la définition des liaisons aromatiques)

- Pour le moment, seules les mesures de similarité/diversité avec les coefficients de Cosinus, Simpson et Tanimoto sont calculées. L'outil a été conçu de façon à ce que d'autres mesures puissent être implémentées avec une perte négligeable de temps de calcul et une difficulté minimale.

- Pour l'instant, seules quatre propriétés physicochimiques sont utilisées: caractère hydrophobe, caractère accepteur de proton, aromaticité et polarisabilité. D'autres propriétés peuvent être implémentées par construction de règles, à partir des informations disponibles dans le fichier XML.

De plus, quelques points positifs peuvent être remarqués :

+ Possibilité d'effectuer différents calculs de similarité 1-N, N-N, N-M.

+ Choix entre plusieurs mesures de similarité (Simpson, Cosinus et Tanimoto) avec l'opportunité d'effectuer des techniques de « fusion de données » pour obtenir des meilleurs résultats.

+ Possibilité de paramétrer l'importance (le poids) des sous-structures et de certaines propriétés physico-chimiques, avec comme but la personnalisation des mesures de similarité et de diversité.

+ L'introduction de requêtes avec des atomes génériques introduit un degré supplémentaire d'assouplissement et de possibilité de calculs pour l'usager.

Conclusion

Dans ce chapitre, nous avons montré les résultats préliminaires de MolDiA obtenus avec différentes bases de données. Des molécules petites ou grandes, simples ou complexes, ainsi qu'un échantillon de drogues et des molécules courantes en chimique pharmaceutique ont été utilisées. Avec ces données, nous avons fait des analyses de similarité et de diversité en prenant en compte des critères divers, tels que la taille des molécules, leur appartenance à un groupe ou encore leurs propriétés physicochimiques. Après analyse, les deux groupes de données utilisées « Zinc » et « Random » ont montré avoir des configurations moléculaires très différentes. « Zinc » est une base diverse et hétérogène, orienté au *drug design*, tandis que « Random » est une base de petites molécules plus courantes en chimie spectroscopique qu'en chimie médicinale. Au total, ce sont 27 analyses différentes qui sont proposées dans l'outil : analyse simple, avec des poids structuraux, avec des poids structure + propriétés, analyse 1-N, N-N, N-M, et trois mesures de similarité. Les résultats de toutes ces combinaisons peuvent être exploités à travers les fichiers XML générés par l'interface graphique. Ces fichiers contiennent toutes les informations utiles dans un format compatible avec la plupart des outils graphiques modernes, facilitant la tache de post-traitement et d'analyse de résultats. L'outil comprend également des feuilles de style qui permettent d'ouvrir les fichiers de résultats et de visualiser d'une manière rapide et facile les données issues du calcul, et les informations concernant les molécules impliquées : la structure 2D ou 3D (si disponible), ainsi que le nombre d'atomes. Les domaines d'application de MolDiA sont nombreux. Même si la sélection et/ou l'organisation des molécules était la vocation principale de l'application, l'interrogation des bases par rapport à une molécule particulière, ainsi que les analyses de diversité au sein d'un même groupe de molécules, en vue de la création/optimisation des bases existantes, sont tout à fait réalisables. Nos applications concernent donc le « Drug design » et la chimie médicinale, où la recherche de molécules ayant une structure ou propriété particulière est souhaitée.

Références

[Meringer2006] Dr. Markus Meringer. Mathematical Department. University of Bayreuth. Germany. http://www.mathe2.uni-bayreuth.de/markus/markus.html

[Holliday2002] Holliday, J.D., Hu, C.Y. and Willett, P., *Grouping of coefficients for the calculation of Inter-molecular similarity and dissimilarity using 2D fragment Bit-Strings,* Comb. Chem. High Throughput Screening, 5 (2002) 155-166.

[ZincWeb] Irwin and Shoichet, ZINC--a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. 2005; 45(1):177-82. http://zinc.docking.org

CONCLUSIONS

ET PERSPECTIVES

V1.1 Conclusions
V1.2 Perspectives
VI.2.1 Perspectives à moyen terme
VI.2.2 Perspectives à long terme

VI. CONCLUSION ET PERSPECTIVES

VI.1 Conclusions

➢ Nous avons présenté un outil de criblage virtuel (MolDiA) reposant sur une nouvelle conception de la diversité qui inclut des informations structurales et des propriétés physicochimiques. Ce nouveau système a comme but de calculer la similarité et la diversité de bases moléculaires.

➢ Le développement du système MolDiA s'articule autour de trois axes principaux : la création de la base de fragments, la génération des vecteurs descripteurs de molécules et le calcul de la mesure de similarité. Parmi les aspects proposés qui nous paraissent les plus importants dans le développement de l'outil, nous pouvons souligner:

* La base de fragments de MolDiA (FragDB): cette base a été créée manuellement, mais elle a été structurée automatiquement. La FragDB est composée à ce jour de 502 fragments cycliques, de 61 fragments acycliques et de 321 règles d'exclusion qui permettent d'éliminer les doublons lors de recherches structurales.

* La création d'un codage du nom de fichier ainsi qu'une hiérarchie des atomes génériques pour notre base de fragments. Ceci permet d'extraire des informations chimiques difficiles à formaliser ainsi que d'effectuer des requêtes avec des molécules contenant des atomes génériques.

* L'utilisation des langages de marquage (XML) pour la structuration, l'exploitation et l'échange des données chimiques complexes. La base de fragments, les vecteurs descripteurs ainsi que les fichiers de résultats sont indexés et structurés en utilisant les technologies XML. Ces fichiers structurent les informations dans un format compatible avec la plupart des outils graphiques modernes, facilitant la tâche de post-traitement et d'analyse des résultats. L'usage de feuilles de style permet également de visualiser d'une manière rapide et facile, les données issues du calcul, et les

informations concernant les molécules impliquées : la structure 2D ou 3D (si disponible), ainsi que le nombre d'atomes. L'exploitation de ces ressources sur le Web est quasi automatique.

* Le système MolDiA construit dynamiquement des vecteurs descripteurs à partir des informations chimiques extraites du fichier MOL des molécules requêtes et test. Le temps de calcul des vecteurs, dépend de la taille, de la complexité et du nombre de molécules à analyser.

* Des optimisations de l'algorithme d'Ullmann pour la comparaison de graphes chimiques sont utilisées pour effectuer les correspondances molécule - base de fragments. L'inclusion de filtres et de règles sous contraintes nous a permit d'affiner les recherches de sous-structures.

* Les informations au sein des vecteurs descripteurs sont organisées afin de tirer le meilleur résultat de leur nature hétérogène : des clefs de recherche, des informations structurales ainsi que des propriétés physicochimiques sont utilisées pour décrire l'information chimique.

* Des niveaux de comparaison différents combinés avec plusieurs mesures de similarité/ diversité sont proposées. La personnalisation du calcul de la similarité et de la diversité est possible également à travers l'usage des poids structuraux ou des poids de propriétés. Au total, ce sont au moins $3^3 = 27$ analyses différentes qui sont proposées dans l'outil.

➢ Nous avons effectué diverses analyses avec des bases moléculaires différentes. Après avoir analysé les résultats obtenus, nous sommes arrivés aux conclusions suivantes :

* Les deux groupes de données utilisées « Zinc » et « Random » ont montré avoir des configurations moléculaires très différentes. « Zinc » est une base diverse et hétérogène, orientée pour le *drug design*, tandis que « Random » est une base de petites molécules plus courante en chimie spectroscopique qu'en chimie médicinale.

* L'analyse des résultats montrent que les bases contenant des molécules de taille assez grandes (entre 20 et 60 atomes) ont des meilleures performances que celles avec des petites molécules

(nombre d'atomes entre 3 et 15). La raison réside dans la nature des descripteurs utilisés. Plus de sous-structures seront incluses dans le vecteur structural, mieux la molécule sera décrite. Les vecteurs descripteurs de petites molécules, où un ou deux fragments non détectés peuvent fausser les résultats, représentent mal la réalité moléculaire.

* Le calcul de la précision, du rappel et de la *F-mesure* pour les vecteurs descripteurs des molécules de la base « Zinc » est satisfaisant. 76,47% des vecteurs contiennent 100% de correspondance entre le vecteur théorique et celui généré par le système. Un bon rapport précision/rappel est constaté pour environ 80% des vecteurs générés.

* Nous avons observé que l'efficacité du calcul de similarité et de diversité est affectée par la mesure (indice, distance, coefficient) qui est employée pour mesurer le degré de similitude ou de dissimilitude entre les paires de structures. Des trois indices étudiés, un seul est très restrictif (Tanimoto) ce qui assure moins de fausses correspondances entre deux molécules dissimilaires.

➢ Le système MolDiA regroupe une série d'avantages et de limitations, inhérentes au modèle et aux descripteurs choisis. Les aspects positifs et les limitations du logiciel, peuvent être résumés en quelques points :

– La base de fragments contient un nombre fixe de sous-structures et est limitée à une taille relativement petite. Ceci est toutefois extensible. L'idéal serait de pouvoir enrichir FragDB dynamiquement avec des sous-structures extraites de l'UserDB.

– Le traitement de molécules est fait strictement sous format MDL .mol bien défini (voir annexe 3). Les informations moléculaires telles qu'elles sont définies dans les fichiers .mol sont strictement respectées lors du calcul (par exemple, la définition des liaisons aromatiques). Si le fichier .mol d'une molécule n'est pas bien défini (erreur dans la structure, absence d'une marque de fin de fichier, etc.) ceci peut empêcher l'outil de traiter la molécule et interrompe le déroulement du calcul.

– Seuls trois mesures de similarité/diversité (Cosinus, Simpson et Tanimoto) sont calculées pour le moment. L'outil a été conçu de façon à ce que d'autres mesures puissent être implémentées avec une perte négligeable de temps de calcul et une difficulté minimale.

– Pour l'instant, quatre propriétés physicochimiques sont utilisées seulement: caractère hydrophobe, caractère accepteur de proton, aromaticité et polarisabilité. D'autres propriétés peuvent être ajoutées par construction de règles, à partir des informations disponibles dans le fichier XML.

+ MolDiA permet d'effectuer des analyses de bases moléculaires très diverses : petites et grandes molécules, structures simples ou complexes. L'outil montre une bonne efficacité dans un rang assez étendu de diversité moléculaire. Ceci va à l'encontre des modèles actuels qui cherchent à circonscrire l'usage d'un outil à un groupe ou une famille de molécules déterminées.

+ Il est possible d'effectuer des requêtes sous MolDiA avec des molécules contenant des atomes génériques. Ceci introduit un degré supplémentaire d'assouplissement et de possibilités de calculs pour l'usager. Les requêtes génériques, permettent également de cibler peu à peu, le rang de molécules désirées.

+ MolDiA a été dessiné pour offrir le choix entre plusieurs mesures de similarité (Simpson, Cosinus et Tanimoto). L'usager peut utiliser une ou plusieurs mesures à la fois. Ceci ouvre des nouvelles perspectives pour effectuer des techniques de « fusion de données » (*data fusion*) pour obtenir des meilleurs résultats.

+ Il est possible également de paramétrer l'importance (le poids) des sous-structures choisies et de certaines propriétés physico-chimiques pour le calcul de la similarité/diversité. Ceci a comme but d'obtenir des mesures de similarité et de diversité adaptées aux besoins de l'usager.

➢ MolDiA compte avec des multiples applications parmi lesquelles nous pouvons citer :

* La comparaison inter moléculaire entre deux molécules isolées (analyse 1-1), une molécule et une base donnée (analyse 1-N), les molécules d'une seule base (analyse N-N), ainsi que entre deux bases moléculaires différentes (analyse N-M).

* L'analyse des indices de similarité/diversité au sein d'une base moléculaire. Ceci permet à l'utilisateur de juger la composition d'une base donnée (taille, diversité et nature des molécules). Les sous-structures le plus courantes, ainsi que le caractère homogène ou hétérogène dans un espace donné peuvent être déterminées.

* La fusion de données (ou *data fusion*). Ceci est une approche récemment étudiée et qui semble prometteuse. Elle consiste à effectuer plusieurs mesures de similarité avec des indices différents, puis de combiner les résultats selon des règles précises. Ces mesures de similarité sont généralement plus efficaces (en termes de recherche des molécules bio-actives) que des rangs basés sur un seul coefficient simple. Ceci est valable sous condition qu'une combinaison appropriée des coefficients soit choisie pour la fusion. Le coût informatique est également négligeable car les valeurs additionnelles de coefficients peuvent être calculées à un coût informatique minime puisque les comparaisons de sous-structures ont été déjà effectuées pour le premier index.

* Les domaines d'application de MolDiA sont très nombreux, car les méthodes pour calculer les similitudes (ou des dissimilitudes) entre des paires, ou de plus grands groupes de molécules, jouent aujourd'hui un rôle important dans beaucoup d'aspects et domaines de la chimie informatique, tels la construction des bibliothèques, la prévision de propriétés, la conception de systèmes de synthèse, le criblage virtuel et l'analyse moléculaire de diversité.

* L'interrogation de bases de données par rapport à une molécule particulière, ainsi que les analyses de diversité au sein d'une même base moléculaire, sont tout à fait réalisables. Nos applications ne concernent donc pas seulement le « *drug design* » et la chimie médicinale, mais tout domaine où la recherche de molécules ayant une structure ou une propriété particulière est souhaitée.

VI.2 Perspectives

VI.2.1 Perspectives à moyen terme

- Implémentation d'un éditeur d'équations pour l'introduction par l'utilisateur des ses propres indices de similarité.

- Implémentation d'un outil « import/export » des formats moléculaires en utilisant des logiciels existants (e.g. Babel).

- Implémentation d'un outil graphique pour l'introduction par l'utilisateur des molécules requêtes ou des molécules test.

- Extension de la base de fragments pour augmenter le champ des fragments détectés ainsi que la diversité moléculaire des résultats.

VI.2.2 Perspectives à long terme

- Conception et implémentation d'un module de prédiction de propriétés physicochimiques (QSAR).

- Extension des fonctionnalités de l'outil pour son utilisation en biologie moléculaire et bioinformatique.

- Recherche de la Similarité/Diversité pour des molécules en 3D.

- Définition de nouvelles règles pour augmenter le nombre de propriétés physicochimiques incluses dans le modèle.

- Extension des structures de données pour interpréter de nouvelles informations (enantiomers, chiralité, volumes, surfaces…)

- Classification et *clustering* des bases de molécules en utilisant des méthodes mathématiques de classification supervisé (e.g. NN, SVM) et non supervisé (e.g. PCA).

ANNEXES

Annexe 1. Manuel d'utilisation du logiciel
Annexe 2. Fichiers XML et structures de données
Annexe 3. Format MOL
Annexe 4. Tableaux de résultats

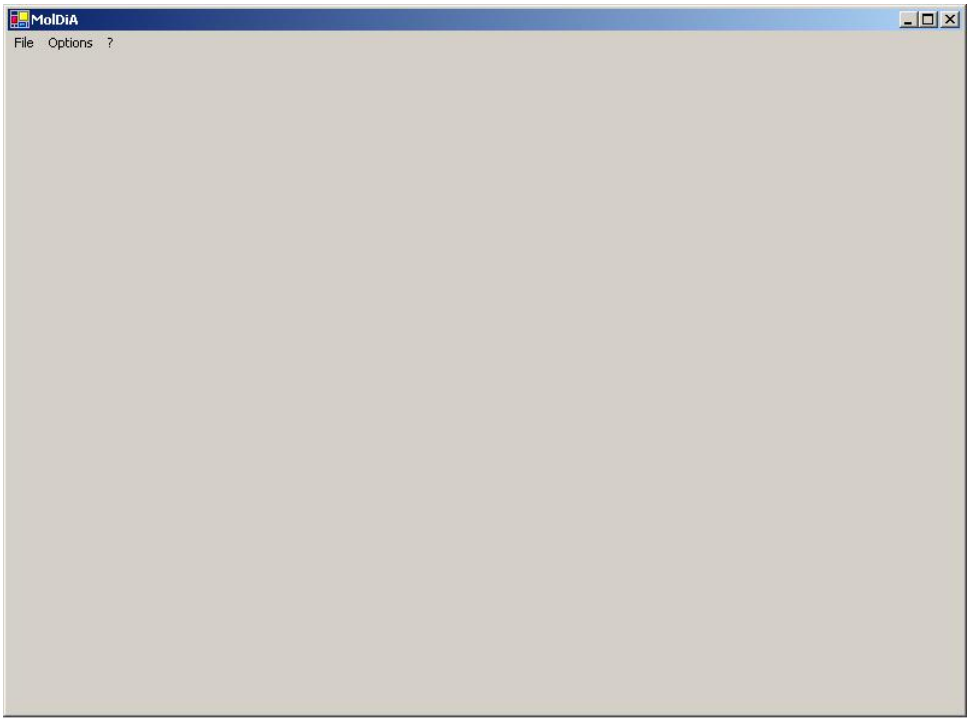
Annexe 1. Manuel d'utilisation du logiciel

Cette annexe a pour but de résumer brièvement l'utilisation du logiciel MolDiA. Le logiciel compte avec une interface simple et claire, qui permet à l'utilisateur d'exploiter les multiples possibilités de calcul offertes. Tout d'abord, un tutoriel pour les usagers débutants avec des étapes simples et claires sera présenté. Ensuite, un module d'analyse rapide pour des usages expérimentés sera abordé. Finalement, on fera une récapitulation des menus et des commandes du logiciel.

A1.1 Tutoriel « Wizard »

L'utilisateur installe le logiciel sur sa machine grâce au « *setup* » généré pour l'occasion.
Au début de l'exécution du logiciel, la FENETRE W0 contient les commandes d'action pour l'outil.

FENETRE W0



Pour accéder au tutoriel débutant, il faut aller au menu « File » puis choisir « Wizard ». La FENETRE W1 est la fenêtre de bienvenue. Elle met en garde l'utilisateur sur l'usage d'autre format que le format MOL et décrit brièvement le but de l'outil.

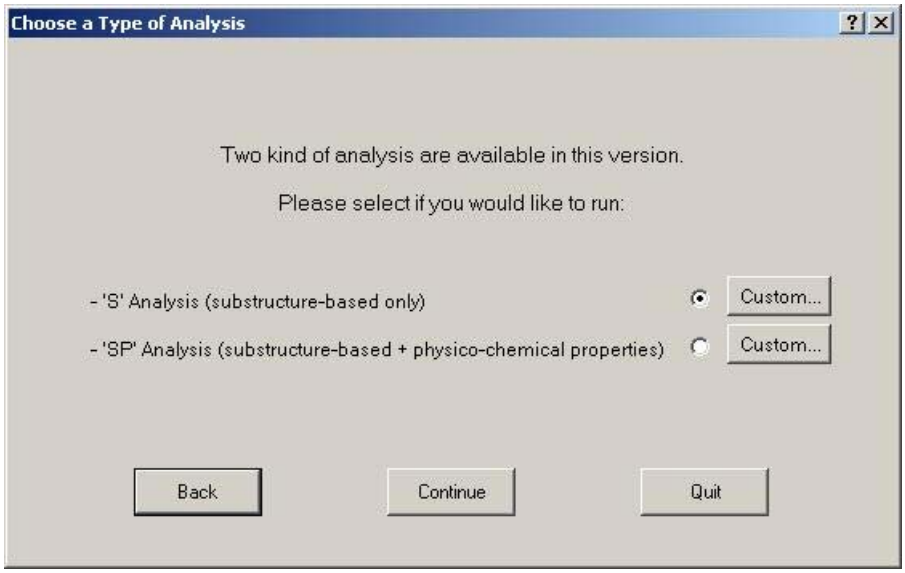
FENETRE W1

Welcome to the MolDiA Wizard
Welcome to the MolDiA Wizard !
This program allows you to make different analysis of similarity and diversity between molecules in a fast and easy way.
/!\ ATTENTION: The molecular format used in this software is MDL .mol; Please, convert your molecules before continue ...
Continue
Quit

Si l'utilisateur choisit QUIT il quitte l'assistant. On peut accéder à l'assistant par le menu File.

Si l'utilisateur choisit CONTINUE la fenêtre suivante apparaît :

FENETRE W2



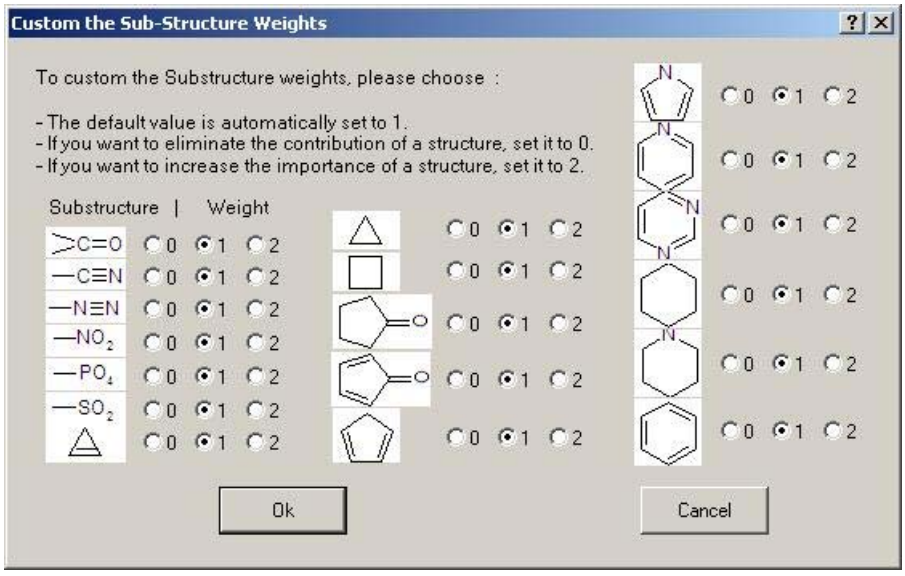
Dans cette fenêtre, on peut choisir soit une analyse dite « S » soit une analyse « SP » :

- L'analyse « S » permet d'effectuer une comparaison des molécules par rapport à leur structure 2D uniquement. Des descripteurs structurels seront construits et l'index de similarité dépendra des sous-structures communes entre la molécule cible et la molécule test.
- L'analyse « SP » permet d'effectuer une comparaison des molécules par rapport à leur structure 2D et les propriétés physicochimiques attachées. Des descripteurs mixtes seront construits et l'index de similarité dépendra des propriétés physicochimiques et des sous-structures communes entre la molécule cible et la molécule test. On peut cocher une seule option (S ou SP) à la fois.

On peut également personnaliser les poids des structures et des propriétés en sélectionnant le bouton CUSTOM… disposé à droite de chaque type d'analyse. Cette sélection fait apparaître une fenêtre qui invite l'utilisateur à paramétrer le calcul selon le cas :

CAS 1) Si on a choisi l'analyse « S » : FENETRE W2-1
CAS 2) Si on a choisi l'analyse « SP » : FENETRE W2-1 puis FENETRE W2-2

FENETRE W2-1

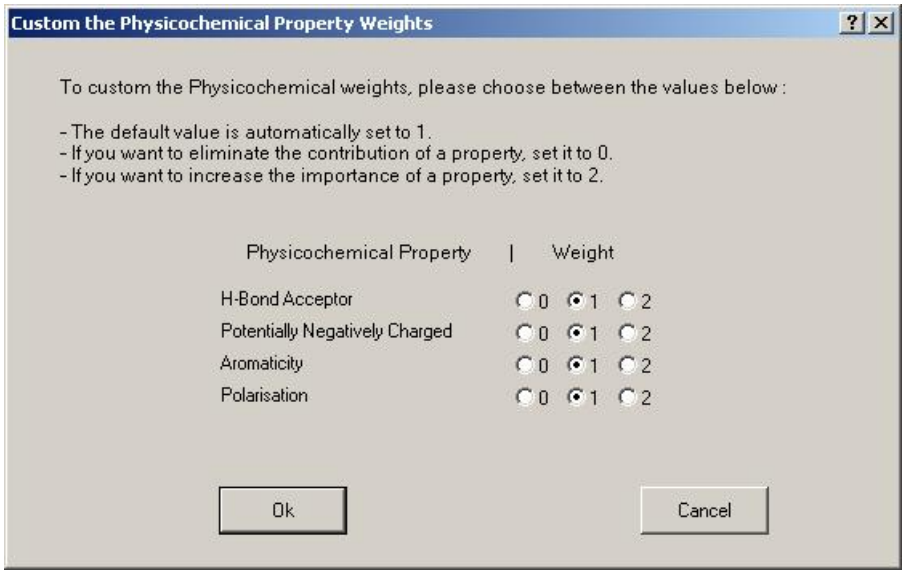

Custom the Sub-Structure Weights
To custom the Substructure weights, please choose :
- The default value is automatically set to 1.
- If you want to eliminate the contribution of a structure, set it to 0.
- If you want to increase the importance of a structure, set it to 2.
Substructure | Weight
>C=O
—C≡N
—N≡N
—NO2
—PO4
—SO2
Ok
Cancel

Dans la FENETRE W2-1, des paramètres de type structurel à personnaliser sont montrés. Les structures sont ordonnées par taille et par complexité : d'abord les groupes acycliques de petites molécules et ensuite les groupes cycliques.

Chaque structure peut prendre trois valeurs différentes :

- Si l'usager ne veut pas changer la valeur par défaut, celle-ci sera 1.
- Si l'usager veut ignorer une ou plusieurs structures de son calcul, il sélectionnera 0.
- Si l'usager veut augmenter la contribution d'une ou plusieurs structures, il sélectionnera 2.

FENETRE W2-2


Custom the Physicochemical Property Weights
To custom the Physicochemical weights, please choose between the values below :
- The default value is automatically set to 1.
- If you want to eliminate the contribution of a property, set it to 0.
- If you want to increase the importance of a property, set it to 2.
Physicochemical Property
Weight
H-Bond Acceptor
Potentially Negatively Charged
Aromaticity
Polarisation
0 1 2
Ok
Cancel

Dans la FENETRE W2-2, les paramètres à personnaliser de type propriété sont montrés : une liste avec des propriétés physicochimiques pertinentes. Comme dans la FENETRE W2-1, cette liste peut être modifiée ou augmentée dans le futur. La valeur par défaut des paramètres est l'unité et les valeurs possibles vont de 0 à 2. La même procédure de la FENETRE W2-1 s'applique ici pour sélectionner ou modifier des poids.

L'utilisateur peut décider de ne pas modifier les valeurs par défaut, donc de ne pas paramétrer le calcul. Dans ce cas il choisit OK dans les FENETRES W2-1 et/ou W2-2 sans rien modifier.

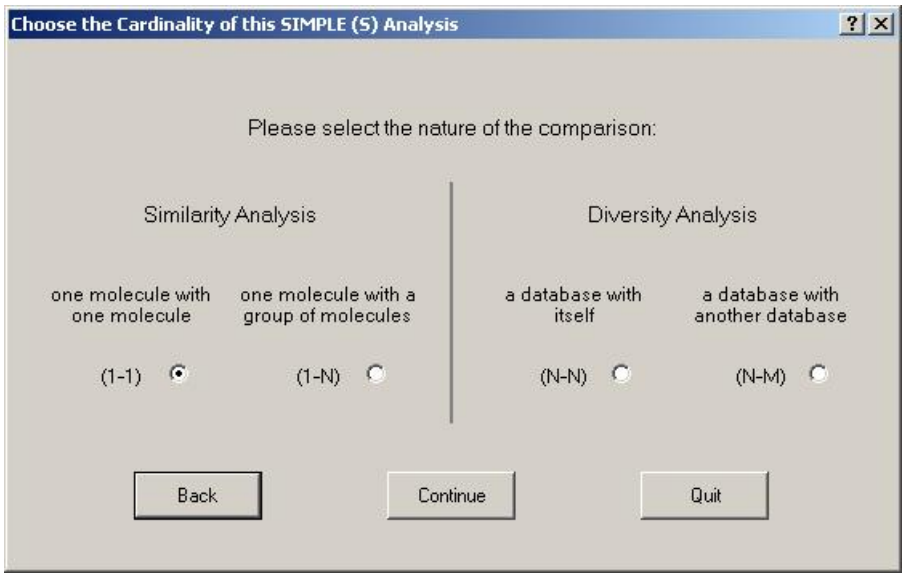
Si l'utilisateur choisit CANCEL on revient en tout cas à la FENETRE W2 qui permet de continuer l'analyse selon le choix fait :

Si l'utilisateur choisit BACK il revient à la FENETRE W1.

Si l'utilisateur choisit QUIT il quitte l'assistant.

Si l'utilisateur choisit CONTINUE la FENETRE W3 apparaît.

FENETRE W3



Dans cette fenêtre, on peut choisir parmi 4 types d'analyse :

- Si l'utilisateur choisit (1-1) : on fera alors un calcul de la similarité entre deux molécules différentes.
- Si l'utilisateur choisit (1-N) : on fera alors un calcul de la similarité entre une molécule et une base.
- Si l'utilisateur choisit (N-N) : on fera alors un calcul de la diversité d'une base de molécules données.
- Si l'utilisateur choisit (N-M) : on fera alors un calcul de la diversité entre deux bases de molécules.

Si l'utilisateur choisit BACK il revient à la FENETRE W2.

Si l'utilisateur choisit QUIT il quitte l'assistant.

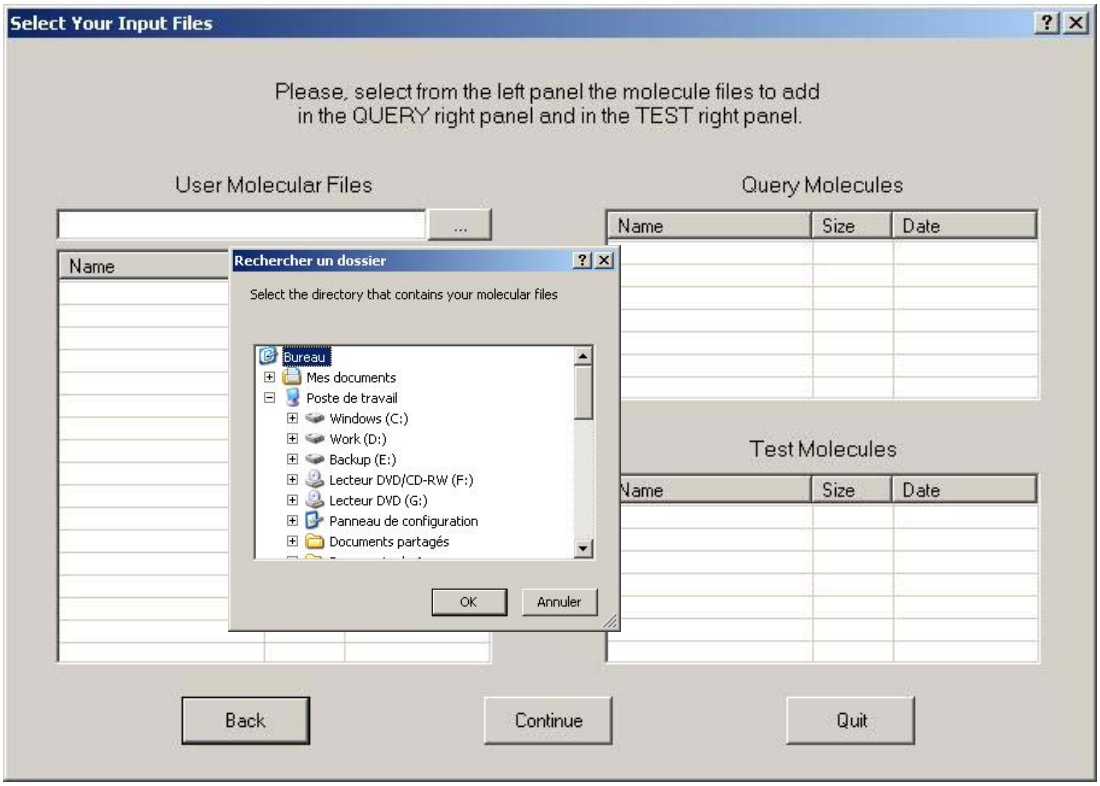
Si l'utilisateur choisit CONTINUE la FENETRE W4 apparaît.

La FENETRE W4 permet de chercher, dans un répertoire déterminé à l'aide du bouton ... (à droite de l'adresse des répertoires) les molécules à analyser.

Dans cette fenêtre on observe trois listes :

La liste de gauche affiche les fichiers avec l'extension .MOL contenus dans le dossier sélectionné. La partie droite contient deux listes. La liste supérieure correspond aux molécules cibles (ou *Query Molécules)* tandis que la liste inférieure sert à indiquer les fichiers des molécules à tester (ou *Test Molécules*).

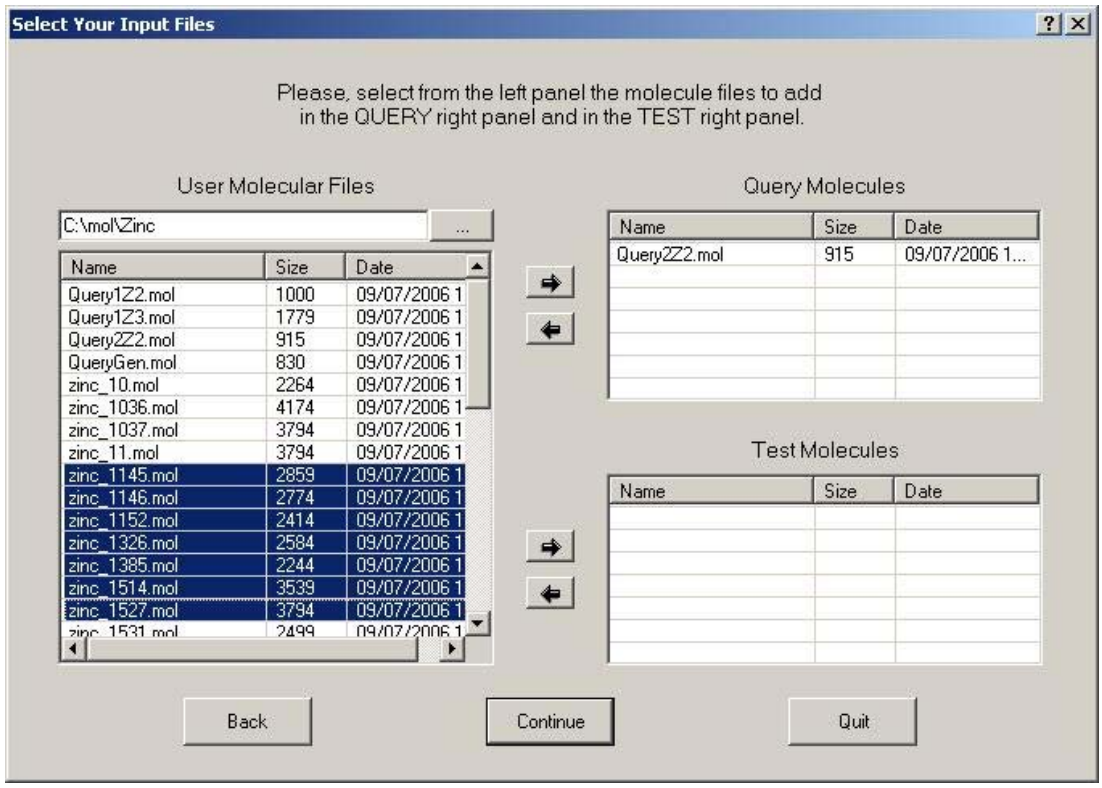
FENETRE W4



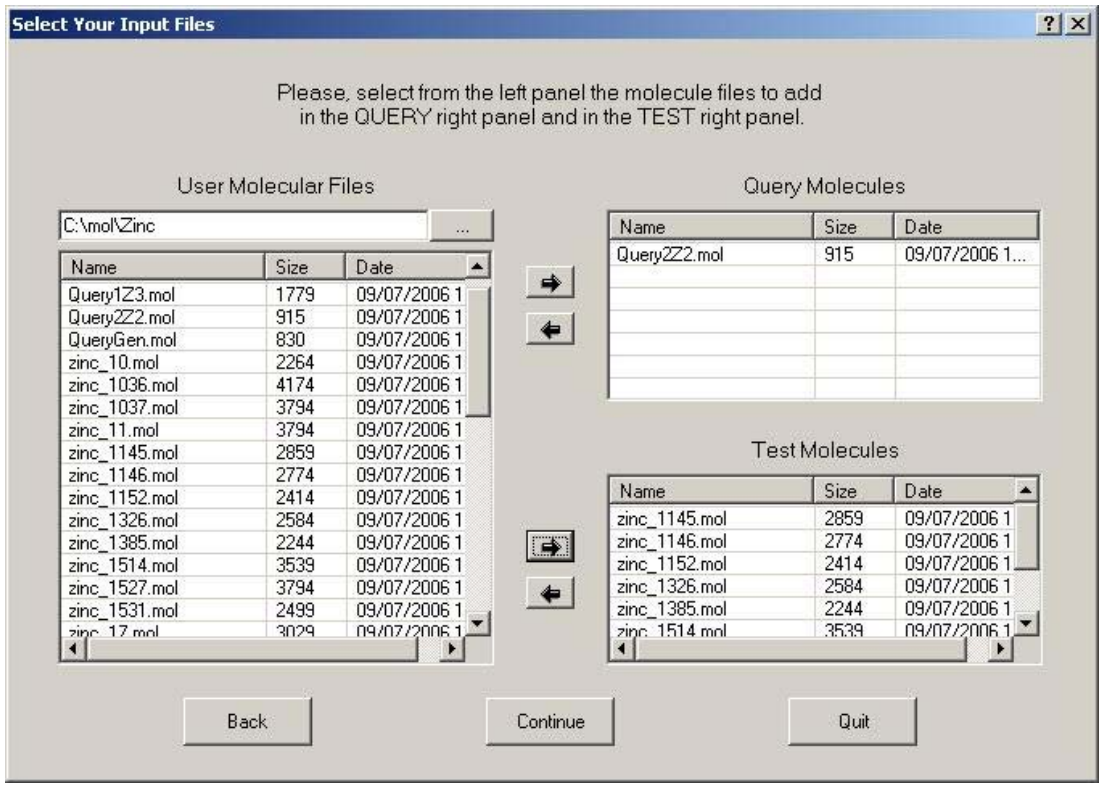
Une fois le répertoire choisi, on peut ajouter ou enlever autant de molécules que l'on souhaite dans les listes *Query* et *Test* à l'aide des flèches → et ← , comme le montre les FENETRES W4-1 et W4-2. De plus, il suffit de garder appuyé les touches « Shift » et « Ctrl » du clavier pour sélectionner un groupe de molécules.

Nous soulignons que le choix des molécules est fait en respectant le type d'analyse (1-1, 1-N, etc.). Ainsi, pour les analyses de type N-N, la fenêtre inférieure droite ne sera pas disponible, car les molécules cibles et tests seront les mêmes. Les molécules pourront donc seulement être entrées dans la fenêtre supérieure droite à l'aide des flèches, de la même manière que expliquée ci-dessus.

FENETRE W4-1



FENETRE W4-2

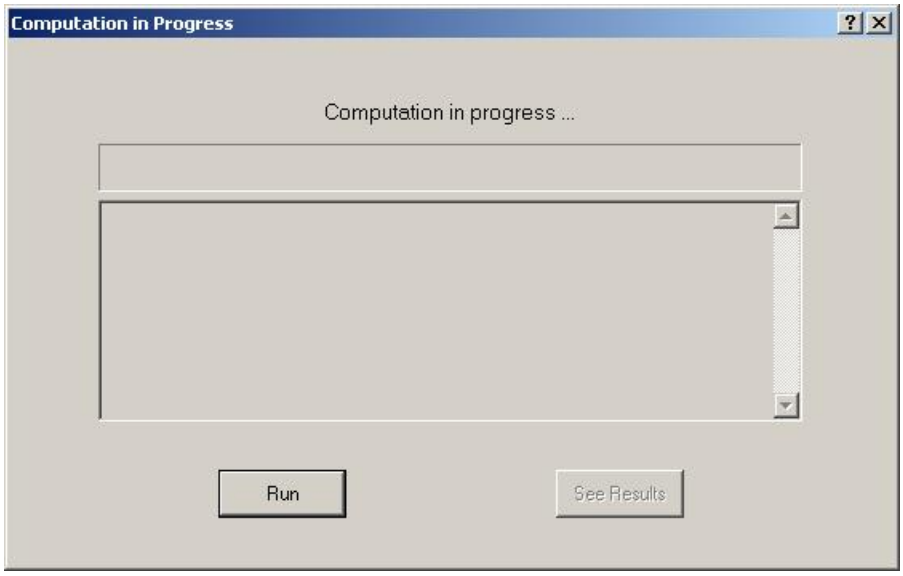


Si l'utilisateur choisit de continuer, le calcul est lancé avec la base de molécules chargées, les valeurs par défaut ou modifiées pour les classes de sous-structures/fragments (cas 1) ou en plus, avec les valeurs par défaut ou modifiées des propriétés physicochimiques (cas 2). La première étape est de charger les informations des molécules dans l'interface. La seconde étape, qui est celle qui prend le plus de temps, est de calculer les vecteurs descripteurs. Une fois que sont calculés les descripteurs, le calcul des indices de similarité/diversité peut être effectué.

Si dans la FENETRE W4 l'utilisateur choisit QUIT il quitte l'assistant. Les données correspondant au calcul seront perdues.
Si l'utilisateur choisit CONTINUE la FENETRE W5 apparaît.

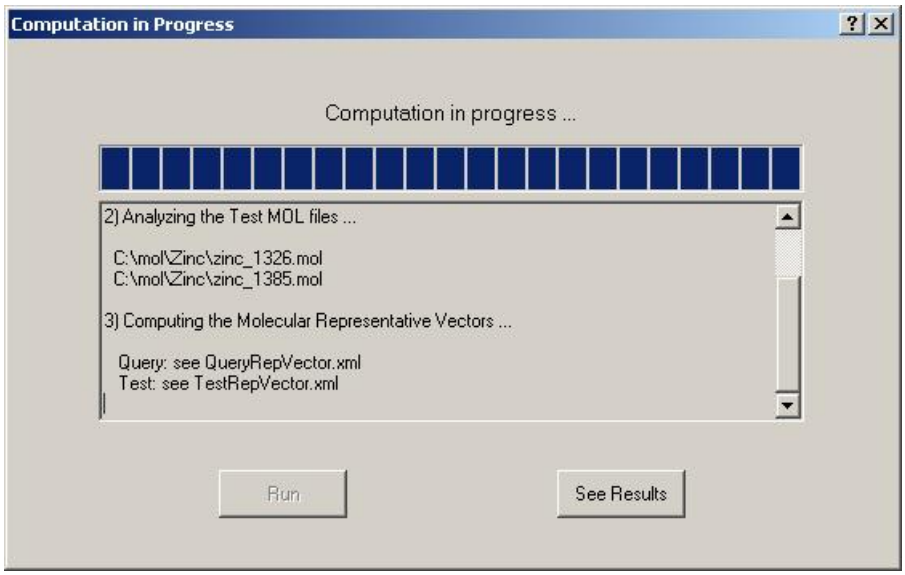
Pour lancer le calcul des vecteurs descripteurs il faut sélectionner la touche RUN de la FENETRE W5. Cette fenêtre montre de manière dynamique le processus de calcul des descripteurs à l'aide d'une barre d'avancement.

FENETRE W5



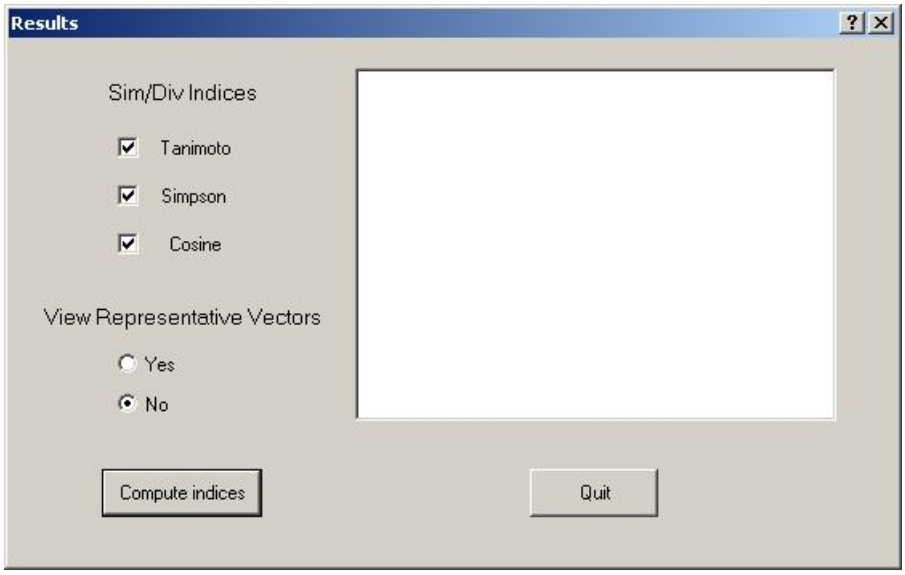
Une fois que le processus de construction des vecteurs est terminé (ce qui peut durer de quelques secondes à plusieurs dizaines de minutes, dépendamment du nombre des molécules, ainsi que de la taille et de la complexité des données), on sélectionne la touche SEE RESULTS pour accéder aux indices de similarité et de diversité.

FENETRE W5-1



Ceci nous amène à la FENETRE W6, dernière étape pour le calcul de la similarité et de la diversité moléculaires. Pour lancer le calcul des indices dans cette fenêtre, il faut sélectionner une ou plusieurs mesures de similarité (Tanimoto, Simpson, Cosinus) puis la touche COMPUTE INDICES

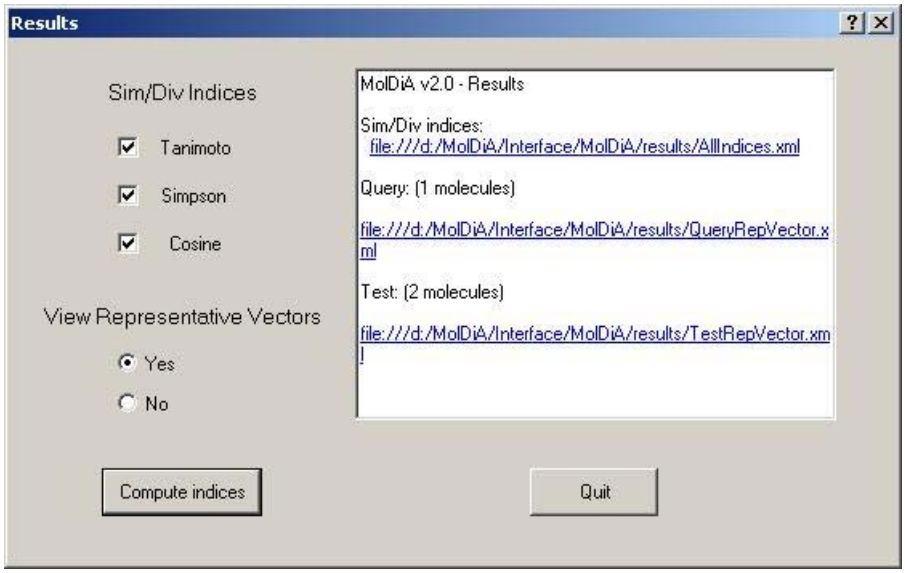
FENETRE W6



On peut également choisir de visualiser ou non, les vecteurs représentatifs calculés dans la FENETRE W5-1. Pour cela il suffit de cocher « Yes » ou « No ».

La FENETRE W6-1 affiche finalement les résultats sous forme d'hyperliens vers les fichiers de résultats en format XML. Ces fichiers seront automatiquement enregistrés dans le répertoire « *results* » du logiciel.

FENETRE W6-1



Les fichiers de résultats peuvent être ouverts en utilisant Internet Explorer pour l'affichage et la visualisation des figures de molécules. On peut également utiliser Excel ou tout autre outil graphique qui supporte le format XML pour effectuer un post traitement sur les tableaux des valeurs (*ranking*, *plotting*, *statistics*, etc.) ainsi que tout outil d'édition de texte pour effectuer des modifications au sein du fichier des résultats.

L'affichage automatique des tableaux de résultats peut être contrôlé en modifiant les feuilles de style XSL placées dans le même dossier que les fichiers de résultats. Le document XSL commande l'apparence que les données auront au moment d'être affichées par Internet Explorer. De la même manière qu'une page web affiche des informations formatées, XML permet de construire des tableaux, des graphes et de personnaliser l'apparence des données issues de MolDiA. Si l'usager désire obtenir les résultats sous forme d'un fichier texte, il suffit de les importer à partir d'un environnement de programmation ou de les éditer à partir de leur affichage sous Internet Explorer.

Pour visualiser correctement les molécules du fichier de résultats sur Internet Explorer il faut s'assurer de l'installation de plugins pour « MDL Chime » (2.6 ou supérieur) et du moteur JAVA pour le script de « JMOL ».

Veuillez noter qu'un déplacement des données moléculaires utilisées pour la génération du fichier résultats (dossier où sont placés les fichiers .MOL), entraîne une perte du lien pour l'affichage des molécules sur les tableaux de résultats. Les plugins d'affichage des molécules « MDL Chime » ou « JMOL » pourront donc générer autant de messages d'erreur que de molécules auront été déplacées. Si le nombre de molécules est élevé, la meilleure manière de fermer Internet Explorer est à travers le « Gestionnaire des taches de Windows ».

Un aperçu des fichiers de résultats ainsi que de vecteurs descripteurs, affichés en utilisant Internet Explorer est donné à continuation :

Figure A1.1. Copies d'écran de fichiers de mesures de similarité N-N ouverts avec Internet Explorer. Les molécules sont présentées en format 3D.

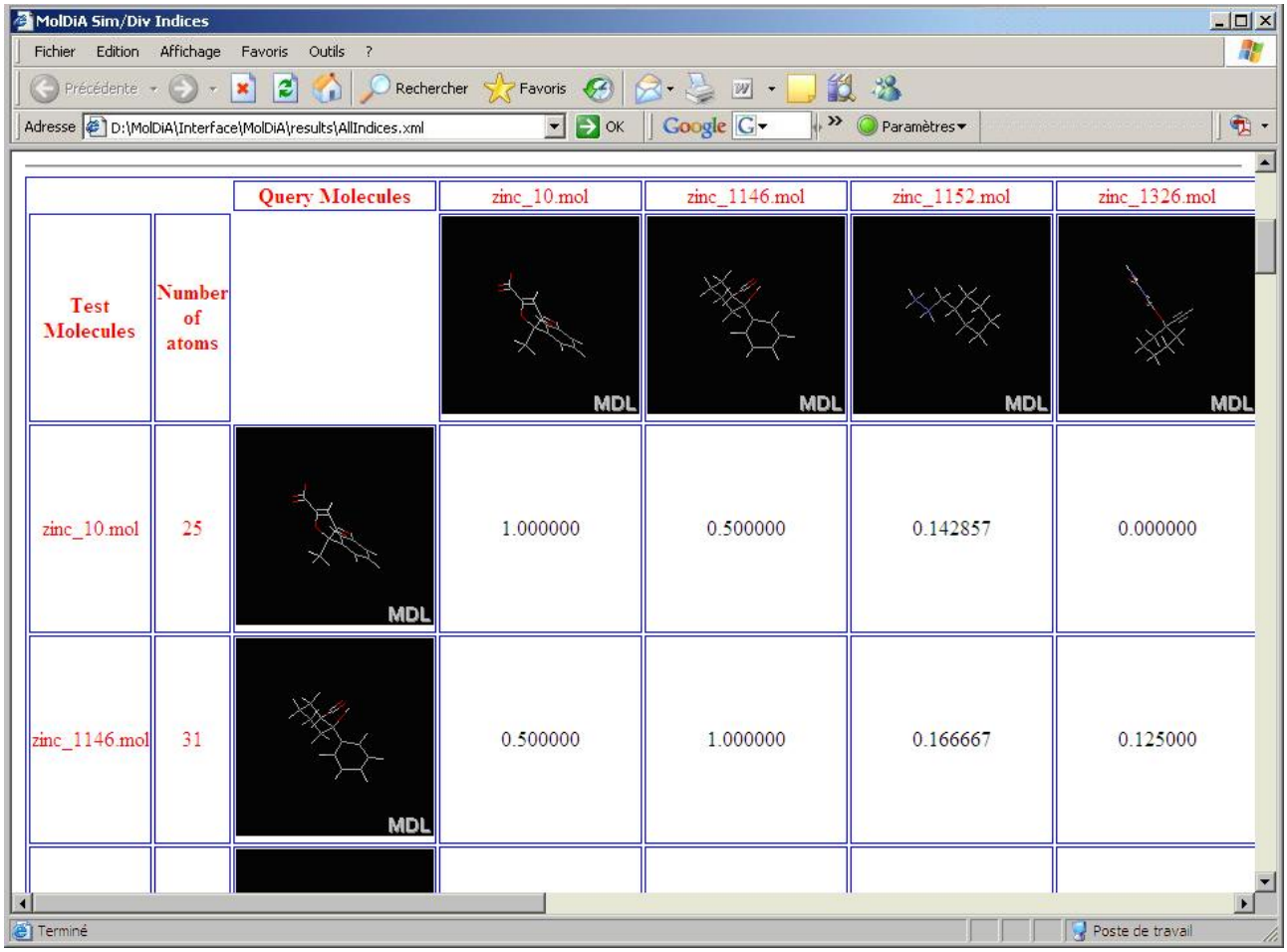
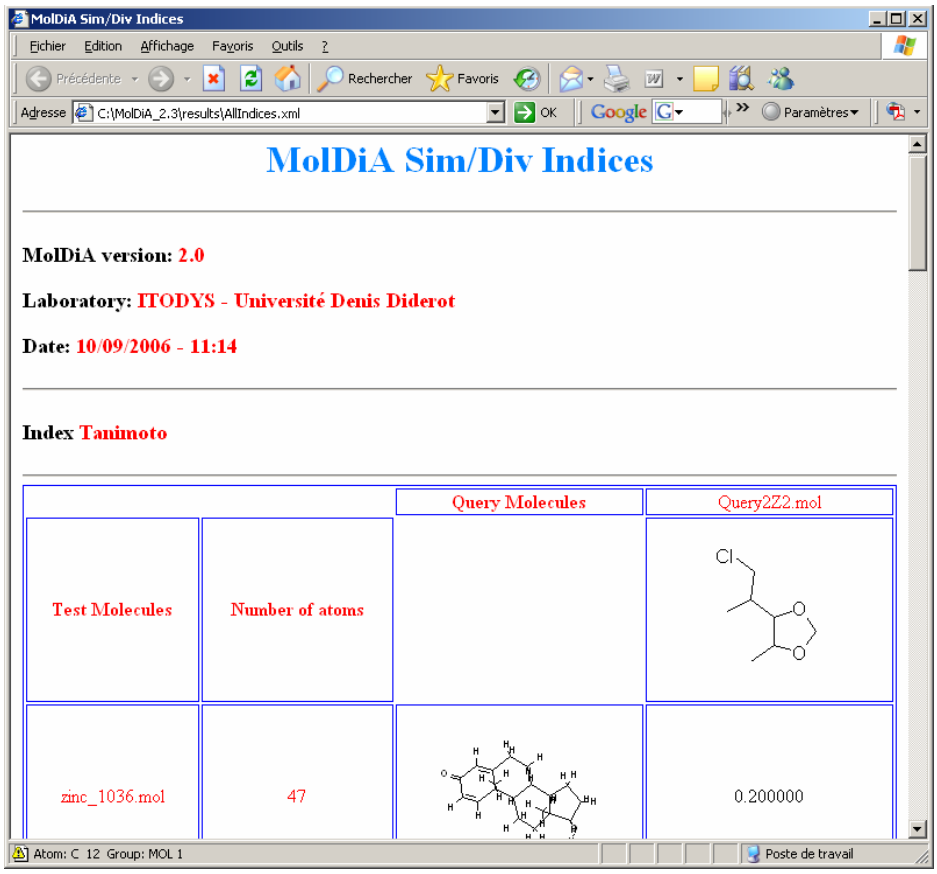


Figure A1.2. Copies d'écran de fichiers de mesures de similarité 1-N ouverts avec Internet Explorer. On observe que les molécules peuvent être présentées en format 2D ou 3D.



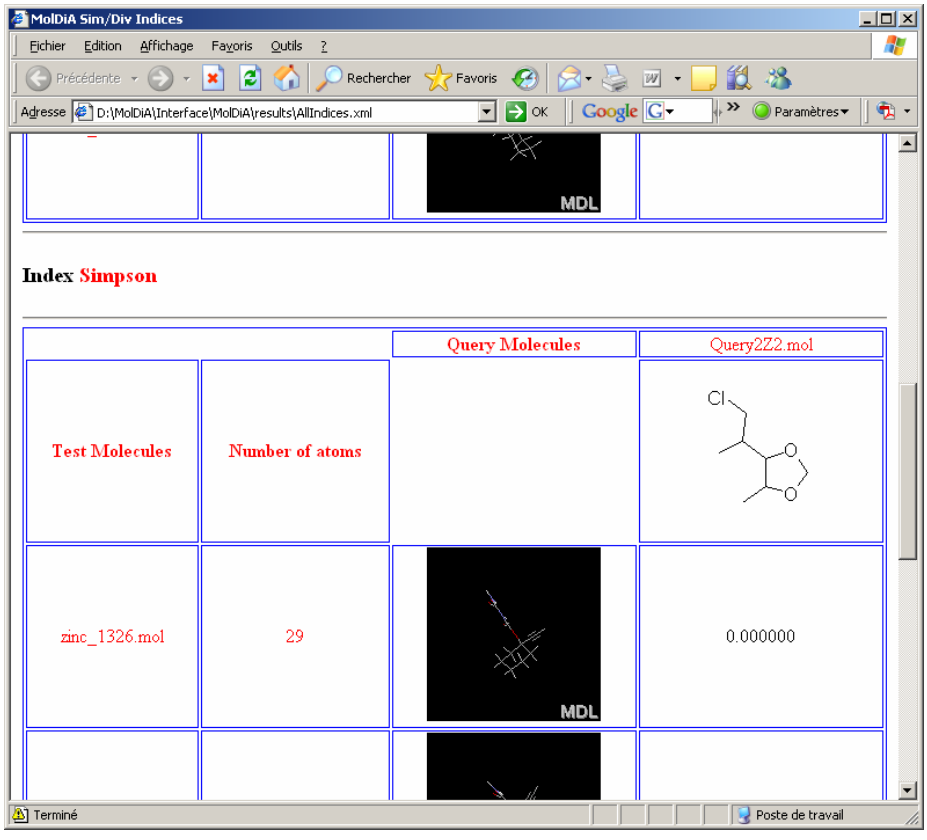
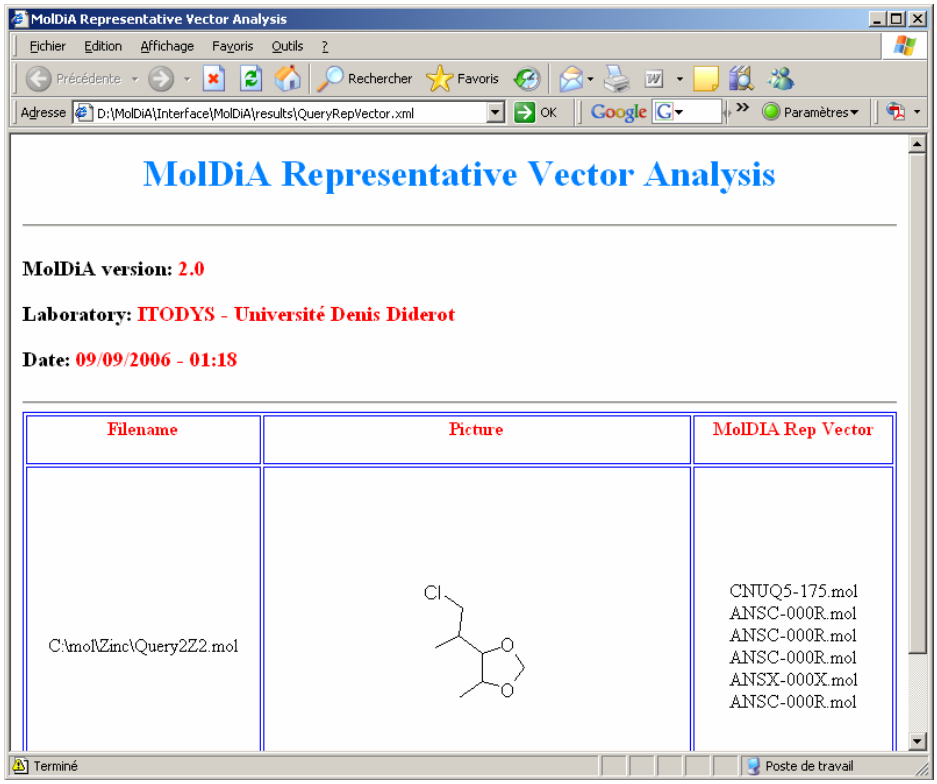
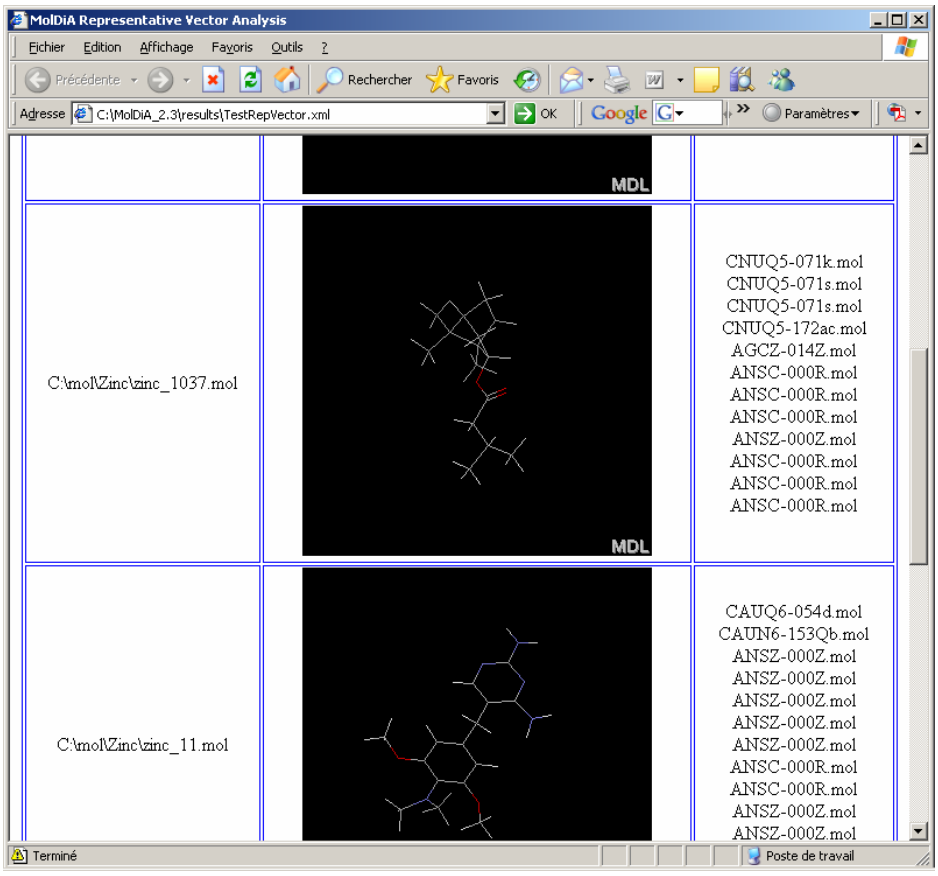


Figure A1.3. Copies d'écran de fichiers de vecteurs résultats ouverts avec Internet Explorer.

Les molécules sont présentées en format 3D et 2D.



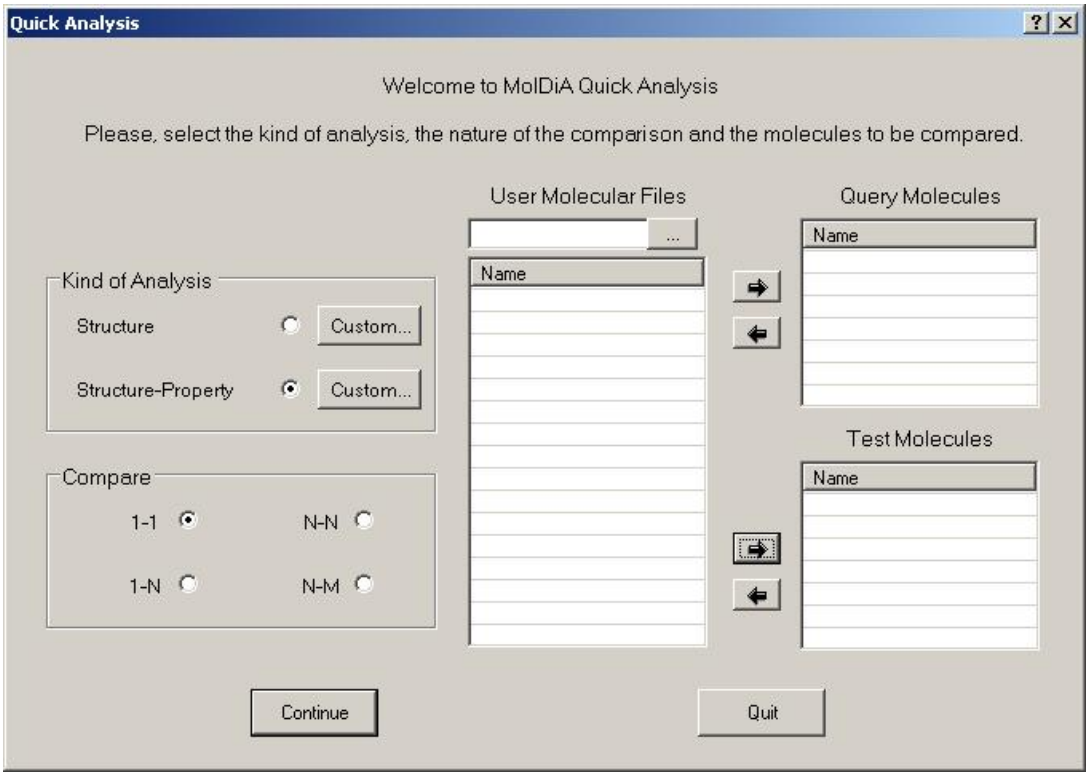


A1.2 Tutoriel « Quick Analysis »

Pour accéder au tutoriel avancé, il faut aller au menu « File » puis choisir « Quick Analysis ».

La FENETRE Q1 montre à l'usager, d'un seul coup d'œil, les types d'analyses et de comparaisons disponibles dans l'outil. Sur la même fenêtre, on peut choisir les molécules à analyser à l'aide du bouton [...]. Les mêmes indications déjà données pour les FENETRES W4, W4-1 et W4-2 s'appliquent à celle-ci.

FENETRE Q1



On peut aussi personnaliser les poids des structures et des propriétés en sélectionnant le bouton [CUSTOM…] disposé à droite de chaque type d'analyse. Cette sélection fait apparaître une fenêtre (FENETRE W2-2 et/ ou FENETRE W2-1) qui invite l'usager à paramétrer le calcul selon le cas.

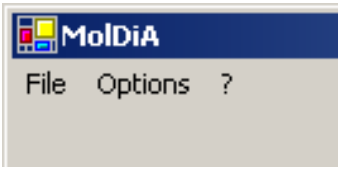
Si l'utilisateur choisit [QUIT] il quitte l'assistant.

Si l'utilisateur choisit [CONTINUE] la FENETRE W5 apparaît.

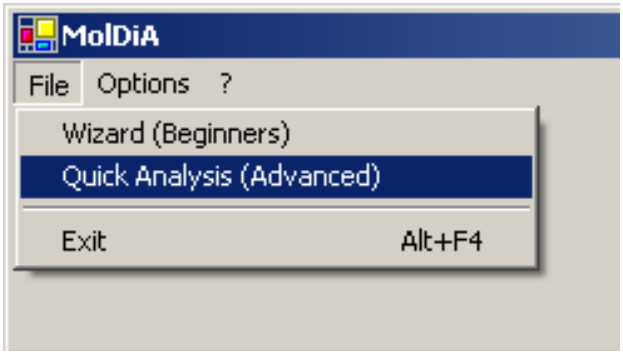
L'usager suit ensuite la même procédure que pour le tutoriel « Wizard », jusqu'à l'obtention des résultats, affichés sur la FENETRE W6-1.

A1.3 Menus du programme

L'interface de MolDIA compte pour l'instant avec les menus suivants : File | Options | Help



Menu File

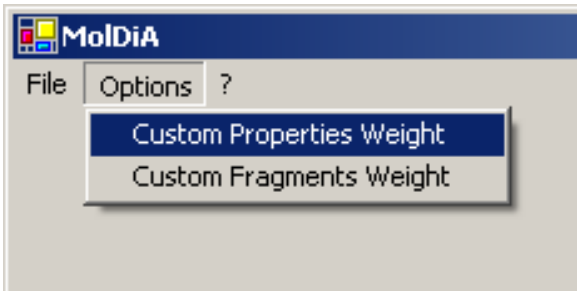


Wizard: accède au tutoriel débutant : FENETRE W1

Quick Analysis : accède au tutoriel avancé FENETRE Q1

Exit : sortie du programme

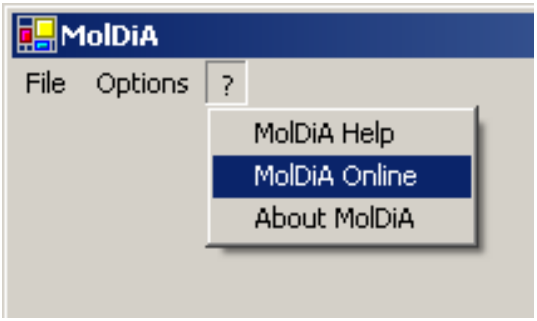
Menu Options



Custom Properties Weight : accède à la FENETRE W2-1

Custom Fragment Weight : accède à la FENETRE W2-2

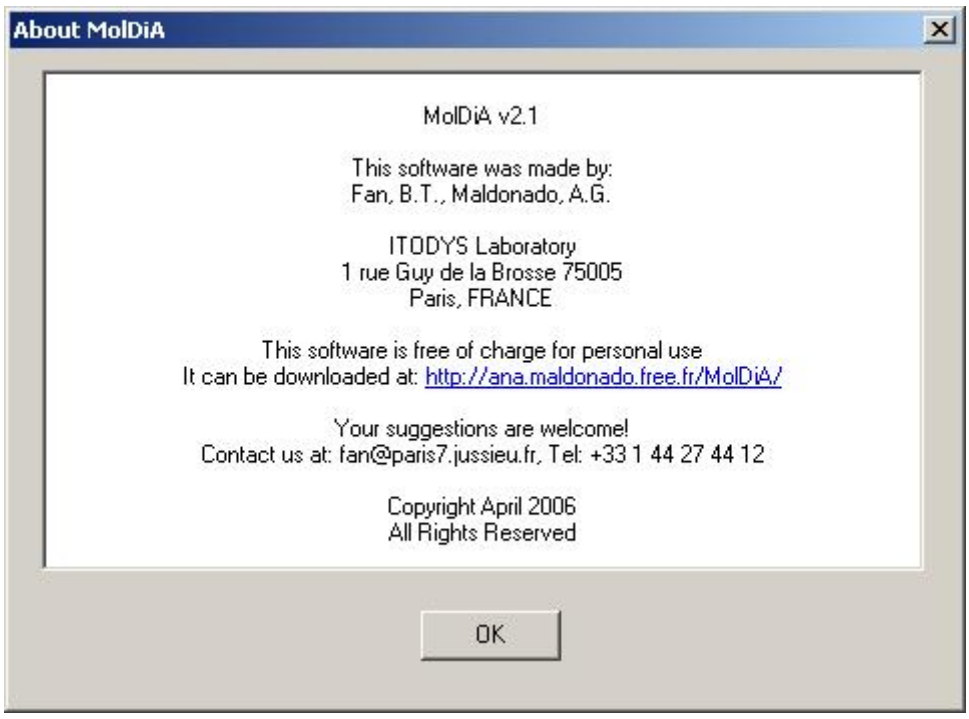
Menu Help



MolDiA Help : accède au présent tutoriel

MolDiA Online: accède à la page web de présentation du logiciel MolDia

About MolDiA : fournit de l'information sur le logiciel MolDiA à travers la fenêtre ci-dessous.


About MolDiA
MolDiA v2.1
This software was made by:
Fan, B.T., Maldonado, A.G.
ITODYS Laboratory
1 rue Guy de la Brosse 75005
Paris, FRANCE
This software is free of charge for personal use
It can be downloaded at: http://ana.maldonado.free.fr/MolDiA/
Your suggestions are welcome!
Contact us at: fan@paris7.jussieu.fr, Tel: +33 1 44 27 44 12
Copyright April 2006
All Rights Reserved
OK

Annexe 2. Fichiers XML et structures de données

A2.1 Structures de données (UserDB et FragDB) en XML et ses DTD

- Modèle de structure de données pour la base de Fragments (FragDB).

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<!-- Model for IndexCyc.xml & IndexAcyc.xml: Data Structure for FragDB  -->
<Index>
      <File name="ici le nom du fichier">
            <Keys>
                  <Key name = "FID" value = "ici code alfa numérique"/>
                  <Key name = "FAtomSum" value = "ici nombre entier"/>
                  <Key name = "FRing" value = "ici valeur binaire"/>
                  <Key name = "FGF" value = "ici chaîne de caractères"/>
            </Keys>
            <Properties>
                  <Property name = "HBondAcceptor" value = "ici nombre entier"/>
                  <Property name = "PotNegCharged" value = "ici nombre entier"/>
                  <Property name = "Aromat" value = "ici nombre entier"/>
                  <Property name = "Polar" value = "ici nombre entier"/>
            </Properties>
      </File>
...

Il y aura autant d'éléments <File></File> comme des fichiers il y a dans FragDB

...

</Index>
```

L'index pour les fragments cycliques (IndexCYC.xml) et acycliques (IndexACYC.xml) de la base de fragments de MolDiA est montré dans la section A2.3

- DTD pour le modèle d'index de fragments : IndexCyc.xml et IndexAcyc.xml

```
<!-- DTD for model of indexCyc and IndexAcyc.xml -->
<!ELEMENT Index (File+)>
<!ELEMENT File (Keys,Properties)>
<!ATTLIST File name CDATA #REQUIRED>
<!ELEMENT Keys (Key+)>
<!ATTLIST Key name CDATA #REQUIRED>
<!ATTLIST Key value CDATA #REQUIRED>
<!ELEMENT Properties (Property+)>
<!ATTLIST Property name CDATA #REQUIRED>
<!ATTLIST Property value CDATA #REQUIRED>
```

- Modèle de la structure de données du vecteur-descripteur pour les molécules de l’usager (UserDB)

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<!-- Model for ResultVector.xml  -->
<indexResultQF>
    <Molecule fileName="ici le nom du fichier à être analysé">
          <ExpRepVector>
              <Frag ref="ici le nom du fragment 1"/>
              ...
              Il y aura autant d’éléments <Frag/> comme des N fragments
              ont été détectes dans la molécule à analyser
              ...

              <Frag ref="ici le nom du fragment N"/>
          </ExpRepVector>
     </Molecule>
...

Il y aura autant d’éléments <Molecule></Molecule> comme des fichiers il y aura à
analyser

...
</indexResultQF>
```

- DTD pour le modèle de structure de données du vecteur-descripteur

```
<!-- DTD for model of ResultVector.xml -->
<!ELEMENT indexResultQF (Molecule+)>
<!ELEMENT Molecule (ExpRepVector)>
<!ATTLIST Molecule fileName CDATA #REQUIRED>
<!ELEMENT ExpRepVector (Frag+)>
<!ATTLIST Frag ref CDATA #REQUIRED >
```

A2.2 Fichier XML des règles d’exclusion des fragments (exclusionRule.xml)

- Fragments Acycliques

```
<?xml version='1.0' encoding='iso-8859-1' standalone='no'?>
<Rule>
<!-- ACYC -->
<FragRule>
  <Cond>
      <Frag ref='AGIS-051R.mol'/>
      <Frag ref='AGIS-051M.mol'/>
      <Frag ref='AGIS-051X.mol'/>
      <Frag ref='AGIS-051Z.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGIN-051Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='ANSC-000R.mol'/>
      <Frag ref='ANSM-000M.mol'/>
      <Frag ref='ANSX-000X.mol'/>
      <Frag ref='ANSZ-000Z.mol'/>
  </Cond>
  <Exc>
      <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCC-014R.mol'/>
      <Frag ref='AGCM-014M.mol'/>
      <Frag ref='AGCX-014X.mol'/>
      <Frag ref='AGCZ-014Z.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGCQ-014Q.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCA-013H.mol'/>
      <Frag ref='AGCA-014A.mol'/>
      <Frag ref='AGCQ-013Q.mol'/>
      <Frag ref='AGCQ-014Q.mol'/>
  </Cond>
  <Exc>
      <Frag ref='ANIZ-002Z.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCQ-014O.mol'/>
  </Cond>
  <Exc>
      <Frag ref='ANIZ-002Z.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
  </Exc>
</FragRule>
<FragRule>
```

```
  <Cond>
      <Frag ref='AGCN-028Q.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGCQ-014Q.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCN-031Q.mol'/>
      <Frag ref='AGCO-015Q.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGCQ-014Q.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCN-030Q.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGCQ-014Q.mol'/>
      <Frag ref='AGCQ-014Q.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCN-031Q.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGCQ-014Q.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCO-015A.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGCA-014A.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCO-016A.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGCA-014A.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
```

```
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCO-016Q.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGCQ-014Q.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGCQ-018Q.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGCQ-014Q.mol'/>
      <Frag ref='AGCQ-014Q.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
      <Frag ref='ANIZ-002Z.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGIN-032Q.mol'/>
  </Cond>
  <Exc>
      <Frag ref='ANIZ-002Z.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGIA-038A.mol'/>
      <Frag ref='AGIE-038R.mol'/>
      <Frag ref='AGIE-053A.mol'/>
      <Frag ref='AGIE-053Q.mol'/>
      <Frag ref='AGIM-038M.mol'/>
      <Frag ref='AGIX-038X.mol'/>
      <Frag ref='AGIZ-038Z.mol'/>
  </Cond>
  <Exc>
      <Frag ref='ANIZ-001Z.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
      <Frag ref='AGIS-051A.mol'/>
  </Cond>
  <Exc>
      <Frag ref='AGIS-051H.mol'/>
  </Exc>
</FragRule>
```

- Fragments Cycliques

```
<!-- CYC -->
<FragRule>
  <Cond>
    <Frag ref='CAUQ3-067.mol'/>
  </Cond>
  <Exc>
    <Frag ref='CNUQ3-132a.mol'/>
    <Frag ref='CNUQ3-132b.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CAUQ3-067a.mol'/>
  </Cond>
  <Exc>
    <Frag ref='CNUQ3-132a.mol'/>
    <Frag ref='CNUQ3-132b.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CAUQ3-067b.mol'/>
  </Cond>
  <Exc>
    <Frag ref='CNUQ3-132c.mol'/>
    <Frag ref='CNUQ3-132d.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CAUQ3-067c.mol'/>
  </Cond>
  <Exc>
    <Frag ref='CNUQ3-132c.mol'/>
    <Frag ref='CNUQ3-132d.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CAUQ3-067d.mol'/>
  </Cond>
  <Exc>
    <Frag ref='CNUQ3-132b.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CAUQ3-067e.mol'/>
  </Cond>
  <Exc>
    <Frag ref='CNUQ3-132d.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CAUQ3-067f.mol'/>
```

```
  </Cond>
  <Exc>
    <Frag ref='CNUQ3-132e.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CNUQ3-131i.mol'/>
    <Frag ref='CNUQ3-131j.mol'/>
    <Frag ref='CNUQ3-132e.mol'/>
    <Frag ref='CNUQ4-195ab.mol'/>
    <Frag ref='CNUQ4-195ac.mol'/>
    <Frag ref='CNUQ5-172bc.mol'/>
    <Frag ref='CNUQ5-172bd.mol'/>
    <Frag ref='CNUQ5-172be.mol'/>
    <Frag ref='CNUQ5-181q.mol'/>
    <Frag ref='CNUQ6-167bv.mol'/>
  </Cond>
  <Exc>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CNUN6-164aa.mol'/>
    <Frag ref='CNUQ3-131g.mol'/>
    <Frag ref='CNUQ3-131h.mol'/>
    <Frag ref='CNUQ4-195aa.mol'/>
    <Frag ref='CNUQ4-195y.mol'/>
    <Frag ref='CNUQ4-195z.mol'/>
    <Frag ref='CNUQ5-172av.mol'/>
    <Frag ref='CNUQ5-172ax.mol'/>
    <Frag ref='CNUQ5-172bb.mol'/>
    <Frag ref='CNUQ5-175.mol'/>
    <Frag ref='CNUQ6-074br.mol'/>
    <Frag ref='CNUQ6-074bs.mol'/>
  </Cond>
  <Exc>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CNUN6-164ad.mol'/>
    <Frag ref='CNUQ3-131e.mol'/>
    <Frag ref='CNUQ4-195q.mol'/>
    <Frag ref='CNUQ4-195r.mol'/>
    <Frag ref='CNUQ4-195u.mol'/>
    <Frag ref='CNUQ4-195v.mol'/>
    <Frag ref='CNUQ4-195w.mol'/>
    <Frag ref='CNUQ5-172ak.mol'/>
    <Frag ref='CNUQ5-172ao.mol'/>
    <Frag ref='CNUQ5-172as.mol'/>
    <Frag ref='CNUQ6-074bl.mol'/>
    <Frag ref='CNUQ6-074bm.mol'/>
    <Frag ref='CNUQ6-074bn.mol'/>
    <Frag ref='CNUQ6-074bo.mol'/>
    <Frag ref='CNUQ6-074bp.mol'/>
  </Cond>
  <Exc>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
```

```
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CNUN6-164ab.mol'/>
    <Frag ref='CNUQ3-131c.mol'/>
    <Frag ref='CNUQ4-195k.mol'/>
    <Frag ref='CNUQ4-195m.mol'/>
    <Frag ref='CNUQ4-195n.mol'/>
    <Frag ref='CNUQ4-195o.mol'/>
    <Frag ref='CNUQ4-195p.mol'/>
    <Frag ref='CNUQ5-172aa.mol'/>
    <Frag ref='CNUQ5-172ac.mol'/>
    <Frag ref='CNUQ5-172af.mol'/>
    <Frag ref='CNUQ5-172ag.mol'/>
    <Frag ref='CNUQ5-174af.mol'/>
    <Frag ref='CNUQ6-074bb.mol'/>
    <Frag ref='CNUQ6-074bc.mol'/>
    <Frag ref='CNUQ6-074bf.mol'/>
    <Frag ref='CNUQ6-074bg.mol'/>
    <Frag ref='CNUQ6-074bh.mol'/>
    <Frag ref='CNUQ6-074bi.mol'/>
    <Frag ref='CNUQ6-074bk.mol'/>
    <Frag ref='CNUQ6-152bd.mol'/>
    <Frag ref='CNUQ6-152bj.mol'/>
  </Cond>
  <Exc>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CNUN6-164a.mol'/>
    <Frag ref='CNUN6-164ac.mol'/>
    <Frag ref='CNUQ4-195h.mol'/>
    <Frag ref='CNUQ4-195i.mol'/>
    <Frag ref='CNUQ5-172o.mol'/>
    <Frag ref='CNUQ5-172r.mol'/>
    <Frag ref='CNUQ5-172s.mol'/>
    <Frag ref='CNUQ5-172u.mol'/>
    <Frag ref='CNUQ6-074ar.mol'/>
    <Frag ref='CNUQ6-074at.mol'/>
    <Frag ref='CNUQ6-074au.mol'/>
    <Frag ref='CNUQ6-074av.mol'/>
    <Frag ref='CNUQ6-074aw.mol'/>
    <Frag ref='CNUQ6-074ax.mol'/>
    <Frag ref='CNUQ6-074ay.mol'/>
    <Frag ref='CNUQ6-074az.mol'/>
  </Cond>
  <Exc>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CNUQ4-195b.mol'/>
```

```
    <Frag ref='CNUQ5-172f.mol'/>
    <Frag ref='CNUQ6-152ad.mol'/>
  </Cond>
  <Exc>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CNUQ6-074.mol'/>
  </Cond>
  <Exc>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
<FragRule>
  <Cond>
    <Frag ref='CNUO5-105a.mol'/>
  </Cond>
  <Exc>
    <Frag ref='ANIZ-002Z.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>

…… Suite des règles CYC dans environ 60 pages ……

<FragRule>
  <Cond>
      <Frag ref='CNUQ6-195bd.mol'/>
  </Cond>
  <Exc>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
    <Frag ref='ANSQ-000Q.mol'/>
  </Exc>
</FragRule>
</Rule>
```

A2.3 Index de fragments et des résultats au format XML

- Index de fragments acycliques (IndexAcyc.xml)

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<index>
      <File name="AGCA-013H.mol">
            <Keys>
                  <Key name="FID" value="013H"/>
                  <Key name="FAtomSum" value="3"/>
                  <Key name="FRing" value="0"/>
                  <Key name="FGF" value="to complete"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="2"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="1"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="AGCA-014A.mol">
            <Keys>
                  <Key name="FID" value="014A"/>
                  <Key name="FAtomSum" value="4"/>
                  <Key name="FRing" value="0"/>
                  <Key name="FGF" value="to complete"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="2"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="1"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="AGCC-014R.mol">
            <Keys>
                  <Key name="FID" value="014R"/>
                  <Key name="FAtomSum" value="4"/>
                  <Key name="FRing" value="0"/>
                  <Key name="FGF" value="to complete"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="2"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="1"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="AGCM-014M.mol">
            <Keys>
                  <Key name="FID" value="014M"/>
                  <Key name="FAtomSum" value="4"/>
                  <Key name="FRing" value="0"/>
                  <Key name="FGF" value="to complete"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="2"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="1"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
```

```
      </File>


      <File name="AGCN-028Q.mol">
            <Keys>
                  <Key name="FID" value="028Q"/>
                  <Key name="FAtomSum" value="8"/>
                  <Key name="FRing" value="0"/>
                  <Key name="FGF" value="to complete"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="2"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="1"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="AGCN-030Q.mol">
            <Keys>
                  <Key name="FID" value="030Q"/>
                  <Key name="FAtomSum" value="8"/>
                  <Key name="FRing" value="0"/>
                  <Key name="FGF" value="to complete"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="2"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="1"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="AGCN-031Q.mol">
            <Keys>
                  <Key name="FID" value="031Q"/>
                  <Key name="FAtomSum" value="8"/>
                  <Key name="FRing" value="0"/>
                  <Key name="FGF" value="to complete"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="2"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="1"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>

…… Suite de l'indexAcyc dans environ 13 pages ……

      <File name="ANSZ-000Z.mol">
            <Keys>
                  <Key name="FID" value="000Z"/>
                  <Key name="FAtomSum" value="2"/>
                  <Key name="FRing" value="0"/>
                  <Key name="FGF" value="none"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="1"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="1"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
</index>
```

- Index de fragments cycliques (IndexCyc.xml)

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<index>
      <File name="CAUA6-055A.mol">
            <Keys>
                  <Key name="FID" value="055A"/>
                  <Key name="FAtomSum" value="12"/>
                  <Key name="FRing" value="6"/>
                  <Key name="FGF" value="none"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="1"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="2"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="CAUA6-055AA.mol">
            <Keys>
                  <Key name="FID" value="055AA"/>
                  <Key name="FAtomSum" value="12"/>
                  <Key name="FRing" value="6"/>
                  <Key name="FGF" value="none"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="1"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="2"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="CAUN5-156Q.mol">
            <Keys>
                  <Key name="FID" value="156Q"/>
                  <Key name="FAtomSum" value="9"/>
                  <Key name="FRing" value="5"/>
                  <Key name="FGF" value="none"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="1"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="2"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="CAUN5-156Qa.mol">
            <Keys>
                  <Key name="FID" value="156Qa"/>
                  <Key name="FAtomSum" value="8"/>
                  <Key name="FRing" value="5"/>
                  <Key name="FGF" value="none"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="1"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="2"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="CAUN5-156Qb.mol">
            <Keys>
                  <Key name="FID" value="156Qb"/>
```

```
                  <Key name="FAtomSum" value="8"/>
                  <Key name="FRing" value="5"/>
                  <Key name="FGF" value="none"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="1"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="2"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="CAUN5-156Qc.mol">
            <Keys>
                  <Key name="FID" value="156Qc"/>
                  <Key name="FAtomSum" value="7"/>
                  <Key name="FRing" value="5"/>
                  <Key name="FGF" value="none"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="1"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="2"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
      <File name="CAUN5-156Qd.mol">
            <Keys>
                  <Key name="FID" value="156Qd"/>
                  <Key name="FAtomSum" value="7"/>
                  <Key name="FRing" value="5"/>
                  <Key name="FGF" value="none"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="1"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="2"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>

…… Suite de l'indexCyc dans environ 122 pages ……

      <File name="CNUQ6-195bd.mol">
            <Keys>
                  <Key name="FID" value="195bd"/>
                  <Key name="FAtomSum" value="11"/>
                  <Key name="FRing" value="6"/>
                  <Key name="FGF" value="none"/>
            </Keys>
            <Properties>
                  <Property name="HBondA" value="1"/>
                  <Property name="PotNCharged" value="1"/>
                  <Property name="Aromat" value="1"/>
                  <Property name="Polar" value="1.0"/>
            </Properties>
      </File>
</index>
```

- Modèle d'index des résultats (indexResult.XML)

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<Query fileName="ici le nom du fichier de la cible">
  <Results>
     <Test fileName="ici le nom du fichier de la molécule test 1">
        <Index Tanimoto= "ici un nombre réel"
          Simpson="ici un nombre réel"
          Cosine="ici un nombre réel"/>
     </Test>
     <Test fileName="ici le nom du fichier de la molécule test 2">
        <Index Tanimoto= "ici un nombre réel"
          Simpson="ici un nombre réel"
          Cosine="ici un nombre réel"/>
     </Test>

      ... suite pour n molecules test ...

     <Test fileName="ici le nom du fichier de la molécule test n">
        <Index Tanimoto= "ici un nombre réel"
          Simpson="ici un nombre réel"
          Cosine="ici un nombre réel"/>
     </Test>
  </Results>
</Query>
```

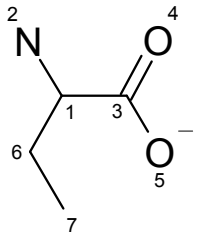
- DTD correspondant au modèle d'index des résultats (indexResult.XML)

```
<!-- Model of DTD for indexResult.XML -->
<!ELEMENT Query (Results)>
<!ATTLIST Query fileName CDATA #REQUIRED>
<!ELEMENT Results (Test+)>
<!ATTLIST Test fileName CDATA #REQUIRED>
<!ELEMENT Test (Index)>
<!ATTLIST Index Tanimoto CDATA #REQUIRED>
<!ATTLIST Index Simpson CDATA #REQUIRED>
<!ATTLIST Index Cosine CDATA #REQUIRED>
```

Annexe 3. Format MOL

Le format du fichier « MOL » était initialement proposé par « *MDL Informations Systems, Inc.* » pour la description numérique des structures moléculaires. Pendant ces dernières années, ce format a été utilisé dans les logiciels appliqués à la chimie, et devient un des formats les plus répandus.

Un fichier MOL est composé principalement par les informations de connectivité, les coordonnées atomiques, des informations associées et la marque à la fin du fichier. On trouve toutes les informations des atomes et des liaisons d'une molécule dans ce fichier. Nous présentons ci-dessous, un exemple dans la figure A3.1




```
  -ISIS-  05170615042D
  7  6  0  0  0  0  0  0  0  0999 V2000
   -0.7291    0.5573    0.0000 C   0  0  3  0  0  0  0  0  0  0  0  0
   -1.1483    1.2609    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    0.0986    0.5668    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.4970    1.2740    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.5198   -0.1459    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
   -1.1284   -0.1539    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7081   -0.8703    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  3  4  2  0  0  0  0
  3  5  1  0  0  0  0
  1  3  1  0  0  0  0
  1  6  1  0  0  0  0
  1  2  1  0  0  0  0
  6  7  1  0  0  0  0
M  CHG  1   5  -1
M  END
```

(a) (b) (c) (d) (e)

a) « Counts Line », b) « Atom Block », c) « Bond Block »,
d) « Stext Block » et «Atom List», e) « Properties Block»

Figure A3.1 Un exemple de fichier MOL.

Dans un fichier MOL, on distingue les sections suivantes:

a) « Counts Line »

Contient principalement les information suivantes : les nombres totaux d'atomes et de liaisons, la liste d'atomes, la marque de chiralité et des informations sur la version. Le format du « Counts Line » est présenté par la suite, et peut être traduit à l'aide du tableau A3.2.

« aaabbblllfffcccsssxxxrrrpppiiimmmvvvvvv »

champ	Description
aaa	Nombre total d'atomes
bbb	Nombre total de liaisons
lll	Nombre total de listes d'atomes
fff	(Abrogé)
ccc	Marque de chiralité, 1= chiral, 0= non chiral
sss	Nombre total de lignes de « Stext »
xxx	Nombre total de composants de réaction + 1
rrr	Nombre total de réactifs
ppp	Nombre total de produits
iii	Nombre total d'espèces intermédiaires
mmm	Champs sans significations dans les nouvelles versions. Valeurs par défaut = 999
vvvvvv	Numero de version de format MOL

Tableau A3.2. Contenu du « Counts Line »

b) « Atom Block »

Contient plusieurs lignes décrivant les informations des atomes. Chaque atome correspond à une ligne de description. Le format d'une ligne de « Atoms Block » est présenté par la suite, et peut être traduit à l'aide du tableau A3.3.

```
« xxxxx.xxxxyyyyy.yyyyzzzzz.zzzz aaaddcccssshhhbbbvvvHHHrrriiimmmnnneee »
```

champ	Description
x y z	Coordonnées des atomes
aaa	Symbole de l'atome
dd	Différence de la masse (masse particulière d'isotope)
ccc	Charge de l'atome
sss	Caractéristique de stéréochimie
hhh	Nombre total d'hydrogènes attachés + 1
bbb	Marque de la caractéristique de stéréochimie considérée
vvv	Valence
HHH	Marque de l'hydrogène attaché
rrr	Type de composant dans la réaction
iii	Nombre total de composants dans la réaction
mmm	Numéro de mappe d'atome-atome
nnn	Marque de changement de configuration
eee	Marque de changement précis dans une réaction

Tableau A3.3. Variables dans « Atom Block »

c) « Bond Block »

Rassemble les informations relatives aux liaisons dans plusieurs lignes. Similaire à « Atom Block », chaque ligne correspond à une liaison. Le format d'une ligne de « Bond Block » est décrit par la suite. Le tableau A3.4 montre la signification des champs.

« 111222tttsssxxxrrrccc »

champ	Description
111222	Numéros d'atomes 1 et 2 de la liaison
ttt	Nature de la liaison : 1= liaison simple, 2= liaison double, 3= liaison triple, 4= liaison aromatiques, 5= liaison simple ou double, 6= liaison simple ou aromatique, 7= liaison double ou aromatiques, 8= quelque soit la nature de la liaison.
sss	Caractéristiques de stéréochimie de liaison. Liaison simple : 1= sortante, 4= sortante ou entrante, 6= entrante. Liaison double : 0= Cis ou Trans déterminé par les coordonnées des atomes, 3= Cis ou Trans
xxx	Non utilisé
rrr	Nature topologique de liaison : 0= incertaine, 1= cycliques, 2= acyclique.
ccc	Situation du centre réactionnel

Tableau A3.4. Champs de « Bond Blocks »

d) « Stext Block» et «Atom List»

Ce champs ne sont pas utilisés dans notre système MolDiA, donc nous ne le détaillons pas ici. Pour plus d'informations on peut consulter la bibliographie.

e) « Properties Block»

Ce bloc regroupe des informations secondaires des atomes, telles que des substituants, des groupes fonctionnels, des super-atomes, la marque de la fin du fichier, etc. Parmi ces informations, celles de caractéristiques secondaires des atomes et la marque de la fin du fichier sont souvent utilisées. Nous présentons dans les paragraphes suivants certaines informations qui sont fréquemment employées.

- **Charge** : Le format de cette information est,

« M CHGnn8 aaa vvv »

Où : « M CHG » est l'indication d'existence de charges. « nn8» est le nombre total d'atomes portant les charges. « aaa » est le numéro d'atome portant la charge, et « vvv » est la valeur de charge portée par cet atome. Si n atomes portent des charges, le format « aaavvv » est répeté n fois. Si aucune charge n'est portée par un atome dan la molécule, cette ligne n'existe pas dans le fichier MOL.

- **Radical Libre** : Le format de ligne enregistrant les informations concernant les radicaux est,

« M RADnn8 aaa vvv »

Où : « M RAD » est l'indication d'existence de radicaux et les autres champs ont les sens analogues que dans la ligne des informations pour les charges. Si aucune charge n'est portée par un atome dan la molécule, cette ligne n'apparaît pas dans le fichier MOL.

- **Isotope** : Le format est,

« M ISOnn8 aaa vvv »

Où : « M ISO » marque la présence d'isotopes dans la molécule. Les autres variables sont similaires à celles pour les charges et les radicaux. Cette ligne sera absente du fichier si la molécule ne contient pas d'isotopes.

- **Marque de la fin du fichier :** Symbolisé par « M END », ceci désigne la fin du fichier.

Références

[Yao2000] Yao, J.H., *Système SIRS-SS: Simulation Spectrale IR et Raman par association sous-structure/sous-spectres*, These de Doctorat. Université Paris 7 - Denis Diderot, Paris, 2000.

[MDLWeb] Information disponible sur: http://www.mdl.com/downloads/public/ctfile/ctfile.pdf

Annexe 4. Tableaux de résultats

A4.1 Résultats d'une analyse « 1-N » avec la base « Zinc »

- Résultats avec la cible Query1Z2.mol

FileName	# atom	Tanimoto	Simpson	Cosine
Query1Z2.mol	11	1,0000	1,0000	1,0000
zinc_10.mol	25	0,6667	0,8000	0,8000
zinc_38.mol	39	0,3636	0,8000	0,5657
zinc_57.mol	48	0,3333	0,8000	0,5394
zinc_58.mol	45	0,3636	0,8000	0,5657
zinc_189.mol	44	0,4000	0,8000	0,5963
zinc_1146.mol	31	0,3333	0,6667	0,5164
zinc_888.mol	37	0,3333	0,6667	0,5164
zinc_7.mol	43	0,2143	0,6000	0,3873
zinc_1037.mol	43	0,2143	0,6000	0,3873
zinc_1527.mol	43	0,2143	0,6000	0,3873
zinc_370.mol	41	0,3000	0,6000	0,4743
zinc_447.mol	45	0,2727	0,6000	0,4472
zinc_32.mol	22	0,1667	0,5000	0,3162
zinc_11.mol	43	0,1250	0,4000	0,2481
zinc_17.mol	34	0,2222	0,4000	0,3651
zinc_18.mol	30	0,2000	0,4000	0,3381
zinc_28.mol	26	0,2500	0,4000	0,4000
zinc_52.mol	35	0,2000	0,4000	0,3381
zinc_53.mol	20	0,2500	0,4000	0,4000
zinc_1036.mol	47	0,2222	0,4000	0,3651
zinc_660.mol	32	0,2222	0,4000	0,3651
zinc_707.mol	36	0,2500	0,4000	0,4000
zinc_948.mol	51	0,2000	0,4000	0,3381
zinc_1145.mol	32	0,1429	0,3333	0,2582
zinc_1152.mol	27	0,1429	0,3333	0,2582
zinc_48.mol	30	0,0909	0,2000	0,1690
zinc_1514.mol	40	0,1111	0,2000	0,2000
zinc_1531.mol	28	0,1111	0,2000	0,2000
zinc_249.mol	38	0,0909	0,2000	0,1690
zinc_252.mol	45	0,1000	0,2000	0,1826
zinc_41.mol	22	0,0000	0,0000	0,0000
zinc_1326.mol	29	0,0000	0,0000	0,0000
zinc_1385.mol	25	0,0000	0,0000	0,0000
zinc_530.mol	34	0,0000	0,0000	0,0000

- Résultats avec la cible Query2Z2.mol

FileName	# atom	Tanimoto	Simpson	Cosine
Query2Z2.mol	10	1,0000	1,0000	1,0000
zinc_18.mol	30	0,4444	0,6667	0,6172
zinc_38.mol	39	0,3333	0,6667	0,5164
zinc_1037.mol	43	0,2857	0,6667	0,4714
zinc_1527.mol	43	0,2857	0,6667	0,4714
zinc_189.mol	44	0,3636	0,6667	0,5443
zinc_447.mol	45	0,3636	0,6667	0,5443
zinc_888.mol	37	0,2857	0,6667	0,4714
zinc_28.mol	26	0,3750	0,6000	0,5477
zinc_707.mol	36	0,3750	0,6000	0,5477
zinc_7.mol	43	0,2000	0,5000	0,3536
zinc_370.mol	41	0,2727	0,5000	0,4330
zinc_660.mol	32	0,3333	0,5000	0,5000
zinc_11.mol	43	0,1176	0,3333	0,2265
zinc_17.mol	34	0,2000	0,3333	0,3333
zinc_57.mol	48	0,1333	0,3333	0,2462
zinc_58.mol	45	0,1429	0,3333	0,2582
zinc_1036.mol	47	0,2000	0,3333	0,3333
zinc_1145.mol	32	0,1250	0,3333	0,2357
zinc_1152.mol	27	0,1250	0,3333	0,2357
zinc_249.mol	38	0,1818	0,3333	0,3086
zinc_948.mol	51	0,1818	0,3333	0,3086
zinc_10.mol	25	0,1000	0,2000	0,1826
zinc_1514.mol	40	0,1000	0,2000	0,1826
zinc_1531.mol	28	0,1000	0,2000	0,1826
zinc_48.mol	30	0,0833	0,1667	0,1543
zinc_252.mol	45	0,0909	0,1667	0,1667
zinc_32.mol	22	0,0000	0,0000	0,0000
zinc_41.mol	22	0,0000	0,0000	0,0000
zinc_52.mol	35	0,0000	0,0000	0,0000
zinc_53.mol	20	0,0000	0,0000	0,0000
zinc_1146.mol	31	0,0000	0,0000	0,0000
zinc_1326.mol	29	0,0000	0,0000	0,0000
zinc_1385.mol	25	0,0000	0,0000	0,0000
zinc_530.mol	34	0,0000	0,0000	0,0000

- Résultats avec la cible Query1Z3.mol

FileName	# atom	Tanimoto	Simpson	Cosine
zinc_530.mol	34	0,0909	1,0000	0,3015
Query1Z3.mol	20	1,0000	1,0000	1,0000
zinc_18.mol	30	0,5000	0,8571	0,6838
zinc_48.mol	30	0,5000	0,8571	0,6838
zinc_38.mol	39	0,6154	0,8000	0,7628
zinc_7.mol	43	0,5333	0,7273	0,6963
zinc_948.mol	51	0,3846	0,7143	0,5698
zinc_17.mol	34	0,3077	0,6667	0,4924
zinc_1036.mol	47	0,3077	0,6667	0,4924
zinc_1152.mol	27	0,1667	0,6667	0,3482
zinc_189.mol	44	0,4286	0,6667	0,6030
zinc_252.mol	45	0,3077	0,6667	0,4924
zinc_447.mol	45	0,4286	0,6667	0,6030
zinc_660.mol	32	0,3077	0,6667	0,4924
zinc_888.mol	37	0,1667	0,6667	0,3482
zinc_370.mol	41	0,3571	0,6250	0,5330
zinc_28.mol	26	0,2308	0,6000	0,4045
zinc_707.mol	36	0,2308	0,6000	0,4045
zinc_11.mol	43	0,3333	0,5455	0,5017
zinc_57.mol	48	0,3750	0,5455	0,5455
zinc_32.mol	22	0,0833	0,5000	0,2132
zinc_58.mol	45	0,3125	0,5000	0,4767
zinc_1037.mol	43	0,2778	0,4545	0,4352
zinc_1527.mol	43	0,2778	0,4545	0,4352
zinc_249.mol	38	0,2000	0,4286	0,3419
zinc_1514.mol	40	0,1429	0,4000	0,2697
zinc_1531.mol	28	0,1429	0,4000	0,2697
zinc_1145.mol	32	0,0769	0,3333	0,1741
zinc_1146.mol	31	0,0769	0,3333	0,1741
zinc_52.mol	35	0,1250	0,2857	0,2279
zinc_1385.mol	25	0,0714	0,2500	0,1508
zinc_10.mol	25	0,0667	0,2000	0,1348
zinc_53.mol	20	0,0667	0,2000	0,1348
zinc_1326.mol	29	0,0667	0,2000	0,1348
zinc_41.mol	22	0,0000	0,0000	0,0000

- Résultats avec la cible QueryGen.mol

FileName	# atom	Tanimoto	Simpson	Cosine
QueryGen.mol	9	1,0000	1,0000	1,0000
zinc_28.mol	26	0,4000	1,0000	0,6325
zinc_249.mol	38	0,2857	1,0000	0,5345
zinc_1145.mol	32	0,2500	0,5000	0,4082
zinc_1152.mol	27	0,2500	0,5000	0,4082
zinc_888.mol	37	0,2500	0,5000	0,4082
zinc_10.mol	25	0,1667	0,5000	0,3162
zinc_1514.mol	40	0,1667	0,5000	0,3162
zinc_1531.mol	28	0,1667	0,5000	0,3162
zinc_707.mol	36	0,1667	0,5000	0,3162
zinc_17.mol	34	0,1429	0,5000	0,2887
zinc_1036.mol	47	0,1429	0,5000	0,2887
zinc_252.mol	45	0,1429	0,5000	0,2887
zinc_660.mol	32	0,1429	0,5000	0,2887
zinc_18.mol	30	0,1250	0,5000	0,2673
zinc_48.mol	30	0,1250	0,5000	0,2673
zinc_948.mol	51	0,1250	0,5000	0,2673
zinc_370.mol	41	0,1111	0,5000	0,2500
zinc_189.mol	44	0,1000	0,5000	0,2357
zinc_447.mol	45	0,1000	0,5000	0,2357
zinc_38.mol	39	0,0909	0,5000	0,2236
zinc_58.mol	45	0,0909	0,5000	0,2236
zinc_57.mol	48	0,0833	0,5000	0,2132
zinc_7.mol	43	0,0769	0,5000	0,2041
zinc_1037.mol	43	0,0769	0,5000	0,2041
zinc_1527.mol	43	0,0769	0,5000	0,2041
zinc_11.mol	43	0,0714	0,5000	0,1961
zinc_32.mol	22	0,0000	0,0000	0,0000
zinc_41.mol	22	0,0000	0,0000	0,0000
zinc_52.mol	35	0,0000	0,0000	0,0000
zinc_53.mol	20	0,0000	0,0000	0,0000
zinc_1146.mol	31	0,0000	0,0000	0,0000
zinc_1326.mol	29	0,0000	0,0000	0,0000
zinc_1385.mol	25	0,0000	0,0000	0,0000
zinc_530.mol	34	0,0000	0,0000	0,0000

A4.2 Résultats d'une analyse « 1-N » avec la base « Random100 »

- Résultats avec la cible RandSel100_16.mol

FileName	# atom	Tanimoto	Simpson	Cosine
RandSel100_16	9	1,0000	1,0000	1,0000
RandSel100_4	7	0,5000	1,0000	0,7071
RandSel100_52	5	0,3333	1,0000	0,5774
RandSel100_6	10	0,7143	0,8333	0,8333
RandSel100_35	8	0,7143	0,8333	0,8333
RandSel100_46	9	0,7143	0,8333	0,8333
RandSel100_89	9	0,6250	0,8333	0,7715
RandSel100_92	9	0,6250	0,8333	0,7715
RandSel100_1	9	0,5714	0,8000	0,7303
RandSel100_15	6	0,5714	0,8000	0,7303
RandSel100_29	10	0,5714	0,8000	0,7303
RandSel100_40	6	0,5714	0,8000	0,7303
RandSel100_67	8	0,5714	0,8000	0,7303
RandSel100_70	7	0,5714	0,8000	0,7303
RandSel100_18	8	0,4286	0,7500	0,6124
RandSel100_63	8	0,4286	0,7500	0,6124
RandSel100_83	8	0,4286	0,7500	0,6124
RandSel100_84	8	0,4286	0,7500	0,6124
RandSel100_86	7	0,4286	0,7500	0,6124
RandSel100_2	9	0,4444	0,6667	0,6172
RandSel100_8	8	0,5000	0,6667	0,6667
RandSel100_25	8	0,2857	0,6667	0,4714
RandSel100_27	9	0,4444	0,6667	0,6172
RandSel100_32	10	0,4444	0,6667	0,6172
RandSel100_37	9	0,4000	0,6667	0,5774
RandSel100_42	7	0,5000	0,6667	0,6667
RandSel100_45	7	0,5000	0,6667	0,6667
RandSel100_47	8	0,4444	0,6667	0,6172
RandSel100_68	7	0,5000	0,6667	0,6667
RandSel100_75	7	0,2857	0,6667	0,4714
RandSel100_77	8	0,4444	0,6667	0,6172
RandSel100_87	8	0,5000	0,6667	0,6667
RandSel100_90	7	0,2857	0,6667	0,4714
RandSel100_95	12	0,3636	0,6667	0,5443
RandSel100_98	12	0,4000	0,6667	0,5774
RandSel100_100	7	0,2857	0,6667	0,4714
RandSel100_22	10	0,3750	0,6000	0,5477
RandSel100_59	6	0,3750	0,6000	0,5477
RandSel100_13	3	0,1429	0,5000	0,2887
RandSel100_20	8	0,1429	0,5000	0,2887
RandSel100_48	6	0,1429	0,5000	0,2887
RandSel100_57	7	0,1429	0,5000	0,2887
RandSel100_58	8	0,2500	0,5000	0,4082
RandSel100_82	8	0,1429	0,5000	0,2887
RandSel100_88	6	0,1429	0,5000	0,2887
RandSel100_96	4	0,1429	0,5000	0,2887

RandSel100_97	6	0,1429	0,5000	0,2887
RandSel100_43	8	0,2222	0,4000	0,3651
RandSel100_61	6	0,2222	0,4000	0,3651
RandSel100_81	6	0,2222	0,4000	0,3651
RandSel100_7	8	0,1250	0,3333	0,2357
RandSel100_24	8	0,2000	0,3333	0,3333
RandSel100_30	7	0,2000	0,3333	0,3333
RandSel100_31	8	0,2000	0,3333	0,3333
RandSel100_36	7	0,1250	0,3333	0,2357
RandSel100_38	8	0,2000	0,3333	0,3333
RandSel100_50	4	0,1250	0,3333	0,2357
RandSel100_76	8	0,2000	0,3333	0,3333
RandSel100_80	5	0,1250	0,3333	0,2357
RandSel100_21	8	0,1111	0,2500	0,2041
RandSel100_41	7	0,1111	0,2500	0,2041
RandSel100_51	7	0,1111	0,2500	0,2041
RandSel100_60	6	0,1111	0,2500	0,2041
RandSel100_72	6	0,1000	0,2000	0,1826
RandSel100_28	9	0,0833	0,1667	0,1543
RandSel100_10	9	0,0000	0,0000	0,0000
RandSel100_11	7	0,0000	0,0000	0,0000
RandSel100_14	8	0,0000	0,0000	0,0000
RandSel100_17	8	0,0000	0,0000	0,0000
RandSel100_19	4	0,0000	0,0000	0,0000
RandSel100_23	6	0,0000	0,0000	0,0000
RandSel100_26	6	0,0000	0,0000	0,0000
RandSel100_62	8	0,0000	0,0000	0,0000
RandSel100_66	3	0,0000	0,0000	0,0000
RandSel100_71	8	0,0000	0,0000	0,0000
RandSel100_74	6	0,0000	0,0000	0,0000
RandSel100_99	8	0,0000	0,0000	0,0000

- Résultats avec la cible RandSel100_29.mol

FileName	# atom	Tanimoto	Simpson	Cosine
RandSel100_29	10	1,0000	1,0000	1,0000
RandSel100_15	6	0,6667	0,8000	0,8000
RandSel100_16	9	0,5714	0,8000	0,7303
RandSel100_2	9	0,5000	0,8000	0,6761
RandSel100_6	10	0,5714	0,8000	0,7303
RandSel100_8	8	0,5714	0,8000	0,7303
RandSel100_27	9	0,5000	0,8000	0,6761
RandSel100_32	10	0,5000	0,8000	0,6761
RandSel100_35	8	0,5714	0,8000	0,7303
RandSel100_37	9	0,4444	0,8000	0,6325
RandSel100_40	6	0,6667	0,8000	0,8000
RandSel100_42	7	0,5714	0,8000	0,7303
RandSel100_45	7	0,5714	0,8000	0,7303
RandSel100_46	9	0,5714	0,8000	0,7303
RandSel100_47	8	0,5000	0,8000	0,6761
RandSel100_68	7	0,5714	0,8000	0,7303
RandSel100_77	8	0,5000	0,8000	0,6761
RandSel100_87	8	0,5714	0,8000	0,7303
RandSel100_89	9	0,5000	0,8000	0,6761
RandSel100_92	9	0,5000	0,8000	0,6761
RandSel100_95	12	0,4000	0,8000	0,5963
RandSel100_98	12	0,4444	0,8000	0,6325
RandSel100_84	8	0,5000	0,7500	0,6708
RandSel100_86	7	0,5000	0,7500	0,6708
RandSel100_4	7	0,3333	0,6667	0,5164
RandSel100_75	7	0,3333	0,6667	0,5164
RandSel100_90	7	0,3333	0,6667	0,5164
RandSel100_1	9	0,4286	0,6000	0,6000
RandSel100_22	10	0,4286	0,6000	0,6000
RandSel100_59	6	0,4286	0,6000	0,6000
RandSel100_67	8	0,4286	0,6000	0,6000
RandSel100_70	7	0,4286	0,6000	0,6000
RandSel100_13	3	0,1667	0,5000	0,3162
RandSel100_18	8	0,2857	0,5000	0,4472
RandSel100_63	8	0,2857	0,5000	0,4472
RandSel100_83	8	0,2857	0,5000	0,4472
RandSel100_88	6	0,1667	0,5000	0,3162
RandSel100_24	8	0,2222	0,4000	0,3651
RandSel100_30	7	0,2222	0,4000	0,3651
RandSel100_31	8	0,2222	0,4000	0,3651
RandSel100_38	8	0,2222	0,4000	0,3651
RandSel100_43	8	0,2500	0,4000	0,4000
RandSel100_61	6	0,2500	0,4000	0,4000
RandSel100_76	8	0,2222	0,4000	0,3651
RandSel100_81	6	0,2500	0,4000	0,4000
RandSel100_25	8	0,1429	0,3333	0,2582
RandSel100_50	4	0,1429	0,3333	0,2582
RandSel100_80	5	0,1429	0,3333	0,2582
RandSel100_21	8	0,1250	0,2500	0,2236
RandSel100_41	7	0,1250	0,2500	0,2236
RandSel100_51	7	0,1250	0,2500	0,2236

RandSel100_58	8	0,1250	0,2500	0,2236
RandSel100_28	9	0,0909	0,2000	0,1690
RandSel100_72	6	0,1111	0,2000	0,2000
RandSel100_10	9	0,0000	0,0000	0,0000
RandSel100_11	7	0,0000	0,0000	0,0000
RandSel100_14	8	0,0000	0,0000	0,0000
RandSel100_17	8	0,0000	0,0000	0,0000
RandSel100_19	4	0,0000	0,0000	0,0000
RandSel100_20	8	0,0000	0,0000	0,0000
RandSel100_7	8	0,0000	0,0000	0,0000
RandSel100_23	6	0,0000	0,0000	0,0000
RandSel100_26	6	0,0000	0,0000	0,0000
RandSel100_36	7	0,0000	0,0000	0,0000
RandSel100_48	6	0,0000	0,0000	0,0000
RandSel100_52	5	0,0000	0,0000	0,0000
RandSel100_57	7	0,0000	0,0000	0,0000
RandSel100_60	6	0,0000	0,0000	0,0000
RandSel100_62	8	0,0000	0,0000	0,0000
RandSel100_66	3	0,0000	0,0000	0,0000
RandSel100_71	8	0,0000	0,0000	0,0000
RandSel100_74	6	0,0000	0,0000	0,0000
RandSel100_82	8	0,0000	0,0000	0,0000
RandSel100_96	4	0,0000	0,0000	0,0000
RandSel100_97	6	0,0000	0,0000	0,0000
RandSel100_99	8	0,0000	0,0000	0,0000
RandSel100_100	7	0,0000	0,0000	0,0000

- Résultats avec la cible RandSel100_51.mol

FileName	#atom	Tanimoto	Simpson	Cosine
RandSel100_13	3	0,5000	1,0000	0,7071
RandSel100_41	7	1,0000	1,0000	1,0000
RandSel100_51	7	1,0000	1,0000	1,0000
RandSel100_88	6	0,5000	1,0000	0,7071
RandSel100_24	8	0,4286	0,7500	0,6124
RandSel100_27	9	0,3750	0,7500	0,5669
RandSel100_28	9	0,3750	0,7500	0,5669
RandSel100_31	8	0,4286	0,7500	0,6124
RandSel100_43	8	0,5000	0,7500	0,6708
RandSel100_76	8	0,4286	0,7500	0,6124
RandSel100_81	6	0,5000	0,7500	0,6708
RandSel100_50	4	0,4000	0,6667	0,5774
RandSel100_80	5	0,4000	0,6667	0,5774
RandSel100_10	9	0,2500	0,5000	0,4082
RandSel100_14	8	0,2000	0,5000	0,3536
RandSel100_15	6	0,2857	0,5000	0,4472
RandSel100_17	8	0,3333	0,5000	0,5000
RandSel100_18	8	0,3333	0,5000	0,5000
RandSel100_19	4	0,2000	0,5000	0,3536
RandSel100_2	9	0,2222	0,5000	0,3780
RandSel100_8	8	0,2500	0,5000	0,4082
RandSel100_21	8	0,3333	0,5000	0,5000
RandSel100_30	7	0,2500	0,5000	0,4082
RandSel100_40	6	0,2857	0,5000	0,4472
RandSel100_42	7	0,2500	0,5000	0,4082
RandSel100_45	7	0,2500	0,5000	0,4082
RandSel100_47	8	0,2222	0,5000	0,3780
RandSel100_59	6	0,2857	0,5000	0,4472
RandSel100_61	6	0,2857	0,5000	0,4472
RandSel100_66	3	0,2000	0,5000	0,3536
RandSel100_67	8	0,2857	0,5000	0,4472
RandSel100_68	7	0,2500	0,5000	0,4082
RandSel100_70	7	0,2857	0,5000	0,4472
RandSel100_72	6	0,2857	0,5000	0,4472
RandSel100_83	8	0,3333	0,5000	0,5000
RandSel100_87	8	0,2500	0,5000	0,4082
RandSel100_92	9	0,2222	0,5000	0,3780
RandSel100_96	4	0,2000	0,5000	0,3536
RandSel100_99	8	0,2000	0,5000	0,3536
RandSel100_4	7	0,1667	0,3333	0,2887
RandSel100_23	6	0,1667	0,3333	0,2887
RandSel100_25	8	0,1667	0,3333	0,2887
RandSel100_26	6	0,1667	0,3333	0,2887
RandSel100_36	7	0,1667	0,3333	0,2887
RandSel100_75	7	0,1667	0,3333	0,2887
RandSel100_90	7	0,1667	0,3333	0,2887
RandSel100_100	7	0,1667	0,3333	0,2887
RandSel100_1	9	0,1250	0,2500	0,2236
RandSel100_11	7	0,1429	0,2500	0,2500
RandSel100_16	9	0,1111	0,2500	0,2041
RandSel100_6	10	0,1111	0,2500	0,2041

RandSel100_22	10	0,1250	0,2500	0,2236
RandSel100_29	10	0,1250	0,2500	0,2236
RandSel100_32	10	0,1000	0,2500	0,1890
RandSel100_35	8	0,1111	0,2500	0,2041
RandSel100_37	9	0,0909	0,2500	0,1768
RandSel100_38	8	0,1111	0,2500	0,2041
RandSel100_46	9	0,1111	0,2500	0,2041
RandSel100_58	8	0,1429	0,2500	0,2500
RandSel100_60	6	0,1429	0,2500	0,2500
RandSel100_62	8	0,1429	0,2500	0,2500
RandSel100_63	8	0,1429	0,2500	0,2500
RandSel100_74	6	0,1429	0,2500	0,2500
RandSel100_77	8	0,1000	0,2500	0,1890
RandSel100_84	8	0,1429	0,2500	0,2500
RandSel100_86	7	0,1429	0,2500	0,2500
RandSel100_89	9	0,1000	0,2500	0,1890
RandSel100_95	12	0,0833	0,2500	0,1667
RandSel100_98	12	0,0909	0,2500	0,1768
RandSel100_20	8	0,0000	0,0000	0,0000
RandSel100_7	8	0,0000	0,0000	0,0000
RandSel100_48	6	0,0000	0,0000	0,0000
RandSel100_52	5	0,0000	0,0000	0,0000
RandSel100_57	7	0,0000	0,0000	0,0000
RandSel100_71	8	0,0000	0,0000	0,0000
RandSel100_82	8	0,0000	0,0000	0,0000
RandSel100_97	6	0,0000	0,0000	0,0000

- Résultats avec la cible RandSel100_74.mol

FileName	#atom	Tanimoto	Simpson	Cosine
RandSel100_74	6	1,000	1,000	1,000
RandSel100_14	8	0,500	1,000	0,707
RandSel100_60	6	0,600	0,750	0,750
RandSel100_13	3	0,200	0,500	0,354
RandSel100_19	4	0,200	0,500	0,354
RandSel100_76	8	0,250	0,500	0,408
RandSel100_88	6	0,200	0,500	0,354
RandSel100_26	6	0,167	0,333	0,289
RandSel100_10	9	0,111	0,250	0,204
RandSel100_15	6	0,125	0,250	0,224
RandSel100_17	8	0,143	0,250	0,250
RandSel100_18	8	0,143	0,250	0,250
RandSel100_41	7	0,143	0,250	0,250
RandSel100_51	7	0,143	0,250	0,250
RandSel100_62	8	0,143	0,250	0,250
RandSel100_81	6	0,125	0,250	0,224
RandSel100_1	9	0,000	0,000	0,000
RandSel100_11	7	0,000	0,000	0,000
RandSel100_16	9	0,000	0,000	0,000
RandSel100_2	9	0,000	0,000	0,000
RandSel100_20	8	0,000	0,000	0,000
RandSel100_4	7	0,000	0,000	0,000
RandSel100_6	10	0,000	0,000	0,000
RandSel100_7	8	0,000	0,000	0,000
RandSel100_8	8	0,000	0,000	0,000
RandSel100_21	8	0,000	0,000	0,000
RandSel100_22	10	0,000	0,000	0,000
RandSel100_23	6	0,000	0,000	0,000
RandSel100_24	8	0,000	0,000	0,000
RandSel100_25	8	0,000	0,000	0,000
RandSel100_27	9	0,000	0,000	0,000
RandSel100_28	9	0,000	0,000	0,000
RandSel100_29	10	0,000	0,000	0,000
RandSel100_30	7	0,000	0,000	0,000
RandSel100_31	8	0,000	0,000	0,000
RandSel100_32	10	0,000	0,000	0,000
RandSel100_35	8	0,000	0,000	0,000
RandSel100_36	7	0,000	0,000	0,000
RandSel100_37	9	0,000	0,000	0,000
RandSel100_38	8	0,000	0,000	0,000
RandSel100_40	6	0,000	0,000	0,000
RandSel100_42	7	0,000	0,000	0,000
RandSel100_43	8	0,000	0,000	0,000
RandSel100_45	7	0,000	0,000	0,000
RandSel100_46	9	0,000	0,000	0,000
RandSel100_47	8	0,000	0,000	0,000
RandSel100_48	6	0,000	0,000	0,000
RandSel100_50	4	0,000	0,000	0,000
RandSel100_52	5	0,000	0,000	0,000
RandSel100_57	7	0,000	0,000	0,000
RandSel100_58	8	0,000	0,000	0,000

RandSel100_59	6	0,000	0,000	0,000
RandSel100_61	6	0,000	0,000	0,000
RandSel100_63	8	0,000	0,000	0,000
RandSel100_66	3	0,000	0,000	0,000
RandSel100_67	8	0,000	0,000	0,000
RandSel100_68	7	0,000	0,000	0,000
RandSel100_70	7	0,000	0,000	0,000
RandSel100_71	8	0,000	0,000	0,000
RandSel100_72	6	0,000	0,000	0,000
RandSel100_75	7	0,000	0,000	0,000
RandSel100_77	8	0,000	0,000	0,000
RandSel100_80	5	0,000	0,000	0,000
RandSel100_82	8	0,000	0,000	0,000
RandSel100_83	8	0,000	0,000	0,000
RandSel100_84	8	0,000	0,000	0,000
RandSel100_86	7	0,000	0,000	0,000
RandSel100_87	8	0,000	0,000	0,000
RandSel100_89	9	0,000	0,000	0,000
RandSel100_90	7	0,000	0,000	0,000
RandSel100_92	9	0,000	0,000	0,000
RandSel100_95	12	0,000	0,000	0,000
RandSel100_96	4	0,000	0,000	0,000
RandSel100_97	6	0,000	0,000	0,000
RandSel100_98	12	0,000	0,000	0,000
RandSel100_99	8	0,000	0,000	0,000
RandSel100_100	7	0,000	0,000	0,000

A4.3 Résultats d’une analyse « N-N » avec la base « Zinc »

- Résultats avec l’index de Tanimoto

- Résultats avec l’index de Simpson

- Résultats avec l’index de Cosinus

TANIMOTO	atom	zinc 10	zinc 1036	zinc 1037	zinc 11	zinc 1145	zinc 1146	zinc 1152	zinc 1326	zinc 1385	zinc 1514	zinc 1527	zinc 1531	zinc 17	zinc 18	zinc 189	zinc 249	zinc 252	zinc 28	zinc 32	zinc 370	zinc 38	zinc 41	zinc 447	zinc 48	zinc 52	zinc 53	zinc 530	zinc 57	zinc 58	zinc 660	zinc 7	zinc 707	zinc 888	zinc 948
zinc 10	25	1,00	0,10	0,13	0,06	0,14	0,50	0,14	0,00	0,00	0,11	0,13	0,11	0,10	0,09	0,27	0,09	0,10	0,11	0,17	0,18	0,25	0,00	0,17	0,09	0,20	0,25	0,00	0,23	0,25	0,10	0,13	0,11	0,14	0,09
zinc 1036	47	0,10	1,00	0,20	0,19	0,13	0,11	0,29	0,10	0,11	0,22	0,20	0,22	0,33	0,30	0,25	0,18	0,20	0,38	0,14	0,27	0,23	0,00	0,25	0,18	0,08	0,10	0,17	0,21	0,23	0,33	0,20	0,22	0,29	0,44
zinc 1037	43	0,13	0,20	1,00	0,14	0,07	0,14	0,15	0,06	0,07	0,13	1,00	0,13	0,20	0,36	0,40	0,12	0,13	0,21	0,08	0,33	0,38	0,00	0,40	0,12	0,12	0,13	0,08	0,21	0,22	0,29	0,26	0,21	0,15	0,19
zinc 11	43	0,06	0,19	0,14	1,00	0,07	0,06	0,14	0,06	0,06	0,13	0,14	0,13	0,27	0,18	0,22	0,18	0,27	0,20	0,07	0,24	0,35	0,00	0,22	0,33	0,11	0,06	0,08	0,33	0,28	0,19	0,47	0,13	0,14	0,18
zinc 1145	32	0,14	0,13	0,07	0,07	1,00	0,00	0,50	0,00	0,00	0,33	0,07	0,14	0,13	0,11	0,09	0,11	0,13	0,14	0,00	0,10	0,08	0,00	0,09	0,11	0,00	0,00	0,00	0,08	0,08	0,29	0,07	0,14	0,20	0,11
zinc 1146	31	0,50	0,11	0,14	0,06	0,00	1,00	0,17	0,13	0,14	0,13	0,14	0,13	0,11	0,10	0,30	0,10	0,11	0,13	0,50	0,20	0,27	0,00	0,18	0,10	0,38	0,50	0,25	0,25	0,27	0,11	0,14	0,00	0,00	0,10
zinc 1152	27	0,14	0,29	0,15	0,14	0,50	0,17	1,00	0,14	0,17	0,60	0,15	0,33	0,29	0,25	0,20	0,25	0,29	0,33	0,25	0,22	0,18	0,00	0,20	0,25	0,11	0,14	0,33	0,17	0,18	0,50	0,15	0,14	0,20	0,25
zinc 1326	29	0,00	0,10	0,06	0,06	0,00	0,13	0,14	1,00	0,80	0,25	0,06	0,67	0,22	0,09	0,08	0,09	0,10	0,11	0,17	0,08	0,07	0,00	0,08	0,09	0,09	0,11	0,20	0,07	0,07	0,22	0,06	0,00	0,00	0,09
zinc 1385	25	0,00	0,11	0,07	0,06	0,00	0,14	0,17	0,80	1,00	0,29	0,07	0,80	0,25	0,10	0,08	0,10	0,11	0,13	0,20	0,09	0,08	0,00	0,08	0,10	0,10	0,13	0,25	0,07	0,08	0,25	0,07	0,00	0,00	0,10
zinc 1514	40	0,11	0,22	0,13	0,13	0,33	0,13	0,60	0,25	0,29	1,00	0,13	0,43	0,38	0,20	0,17	0,20	0,22	0,25	0,17	0,18	0,15	0,00	0,17	0,20	0,09	0,11	0,20	0,14	0,15	0,57	0,13	0,11	0,14	0,20
zinc 1527	43	0,13	0,20	1,00	0,14	0,07	0,14	0,15	0,06	0,07	0,13	1,00	0,13	0,20	0,36	0,40	0,12	0,13	0,21	0,08	0,33	0,38	0,00	0,40	0,12	0,12	0,13	0,08	0,21	0,22	0,29	0,26	0,21	0,15	0,19
zinc 1531	28	0,11	0,22	0,13	0,13	0,14	0,13	0,33	0,67	0,80	0,43	0,13	1,00	0,38	0,20	0,17	0,20	0,22	0,25	0,17	0,18	0,15	0,00	0,17	0,20	0,09	0,11	0,20	0,14	0,15	0,38	0,13	0,11	0,14	0,20
zinc 17	34	0,10	0,33	0,20	0,27	0,13	0,11	0,29	0,22	0,25	0,38	0,20	0,38	1,00	0,30	0,36	0,30	0,33	0,38	0,14	0,40	0,33	0,00	0,36	0,30	0,18	0,10	0,17	0,31	0,33	0,50	0,29	0,22	0,29	0,30
zinc 18	30	0,09	0,30	0,36	0,18	0,11	0,10	0,25	0,09	0,10	0,20	0,36	0,20	0,30	1,00	0,45	0,17	0,18	0,33	0,13	0,36	0,42	0,00	0,45	0,27	0,08	0,09	0,14	0,20	0,21	0,44	0,36	0,33	0,25	0,27
zinc 189	44	0,27	0,25	0,40	0,22	0,09	0,30	0,20	0,08	0,08	0,17	0,40	0,17	0,36	0,45	1,00	0,23	0,25	0,27	0,22	0,70	0,73	0,00	0,64	0,23	0,23	0,17	0,11	0,43	0,46	0,36	0,40	0,27	0,20	0,23
zinc 249	38	0,09	0,18	0,12	0,18	0,11	0,10	0,25	0,09	0,10	0,20	0,12	0,20	0,30	0,17	0,23	1,00	0,30	0,33	0,13	0,25	0,21	0,00	0,23	0,27	0,17	0,09	0,14	0,20	0,21	0,18	0,19	0,09	0,11	0,17
zinc 252	45	0,10	0,20	0,13	0,27	0,13	0,11	0,29	0,10	0,11	0,22	0,13	0,22	0,33	0,18	0,25	0,30	1,00	0,38	0,14	0,27	0,33	0,00	0,25	0,44	0,18	0,10	0,17	0,31	0,33	0,20	0,38	0,10	0,13	0,18
zinc 28	26	0,11	0,38	0,21	0,20	0,14	0,13	0,33	0,11	0,13	0,25	0,21	0,25	0,38	0,33	0,27	0,33	0,38	1,00	0,17	0,30	0,25	0,00	0,27	0,20	0,09	0,11	0,20	0,23	0,25	0,38	0,31	0,25	0,33	0,33
zinc 32	22	0,17	0,14	0,08	0,07	0,00	0,50	0,25	0,17	0,20	0,17	0,08	0,17	0,14	0,13	0,22	0,13	0,14	0,17	1,00	0,11	0,20	0,00	0,10	0,13	0,13	0,17	0,50	0,18	0,20	0,14	0,08	0,00	0,00	0,13
zinc 370	41	0,18	0,27	0,33	0,24	0,10	0,20	0,22	0,08	0,09	0,18	0,33	0,18	0,40	0,36	0,70	0,25	0,27	0,30	0,11	1,00	0,50	0,00	0,70	0,25	0,25	0,30	0,13	0,36	0,38	0,40	0,43	0,30	0,22	0,25
zinc 38	39	0,25	0,23	0,38	0,35	0,08	0,27	0,18	0,07	0,08	0,15	0,38	0,15	0,33	0,42	0,73	0,21	0,33	0,25	0,20	0,50	1,00	0,00	0,58	0,42	0,21	0,15	0,10	0,62	0,54	0,33	0,57	0,25	0,18	0,21
zinc 41	22	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
zinc 447	45	0,17	0,25	0,40	0,22	0,09	0,18	0,20	0,08	0,08	0,17	0,40	0,17	0,36	0,45	0,64	0,23	0,25	0,27	0,10	0,70	0,58	0,00	1,00	0,23	0,23	0,27	0,11	0,33	0,36	0,36	0,40	0,27	0,20	0,23
zinc 48	30	0,09	0,18	0,12	0,33	0,11	0,10	0,25	0,09	0,10	0,20	0,12	0,20	0,30	0,27	0,23	0,27	0,44	0,20	0,13	0,25	0,42	0,00	0,23	1,00	0,17	0,09	0,14	0,38	0,31	0,18	0,46	0,09	0,11	0,17
zinc 52	35	0,20	0,08	0,12	0,11	0,00	0,38	0,11	0,09	0,10	0,09	0,12	0,09	0,18	0,08	0,23	0,17	0,18	0,09	0,13	0,25	0,21	0,00	0,23	0,17	1,00	0,33	0,14	0,20	0,21	0,08	0,19	0,00	0,00	0,08
zinc 53	20	0,25	0,10	0,13	0,06	0,00	0,50	0,14	0,11	0,13	0,11	0,13	0,11	0,10	0,09	0,17	0,09	0,10	0,11	0,17	0,30	0,15	0,00	0,27	0,09	0,33	1,00	0,20	0,14	0,15	0,10	0,13	0,00	0,00	0,09
zinc 530	34	0,00	0,17	0,08	0,08	0,00	0,25	0,33	0,20	0,25	0,20	0,08	0,20	0,17	0,14	0,11	0,14	0,17	0,20	0,50	0,13	0,10	0,00	0,11	0,14	0,14	0,20	1,00	0,09	0,10	0,17	0,08	0,00	0,00	0,14
zinc 57	48	0,23	0,21	0,21	0,33	0,08	0,25	0,17	0,07	0,07	0,14	0,21	0,14	0,31	0,20	0,43	0,20	0,31	0,23	0,18	0,36	0,62	0,00	0,33	0,38	0,20	0,14	0,09	1,00	0,91	0,21	0,44	0,14	0,17	0,20
zinc 58	45	0,25	0,23	0,22	0,28	0,08	0,27	0,18	0,07	0,08	0,15	0,22	0,15	0,33	0,21	0,46	0,21	0,33	0,25	0,20	0,38	0,54	0,00	0,36	0,31	0,21	0,15	0,10	0,91	1,00	0,23	0,38	0,15	0,18	0,21
zinc 660	32	0,10	0,33	0,29	0,19	0,29	0,11	0,50	0,22	0,25	0,57	0,29	0,38	0,50	0,44	0,36	0,18	0,20	0,38	0,14	0,40	0,33	0,00	0,36	0,18	0,08	0,10	0,17	0,21	0,23	1,00	0,29	0,38	0,29	0,30
zinc 7	43	0,13	0,20	0,26	0,47	0,07	0,14	0,15	0,06	0,07	0,13	0,26	0,13	0,29	0,36	0,40	0,19	0,38	0,31	0,08	0,43	0,57	0,00	0,40	0,46	0,19	0,13	0,08	0,44	0,38	0,29	1,00	0,21	0,15	0,19
zinc 707	36	0,11	0,22	0,21	0,13	0,14	0,00	0,14	0,00	0,00	0,11	0,21	0,11	0,22	0,33	0,27	0,09	0,10	0,25	0,00	0,30	0,25	0,00	0,27	0,09	0,00	0,00	0,00	0,14	0,15	0,38	0,21	1,00	0,33	0,20
zinc 888	37	0,14	0,29	0,15	0,14	0,20	0,00	0,20	0,00	0,00	0,14	0,15	0,14	0,29	0,25	0,20	0,11	0,13	0,33	0,00	0,22	0,18	0,00	0,20	0,11	0,00	0,00	0,00	0,17	0,18	0,29	0,15	0,33	1,00	0,25
zinc 948	51	0,09	0,44	0,19	0,18	0,11	0,10	0,25	0,09	0,10	0,20	0,19	0,20	0,30	0,27	0,23	0,17	0,18	0,33	0,13	0,25	0,21	0,00	0,23	0,17	0,08	0,09	0,14	0,20	0,21	0,30	0,19	0,20	0,25	1,00

SIMPSON	atom	zinc 10	zinc 1036	zinc 1037	zinc 11	zinc 1145	zinc 1146	zinc 1152	zinc 1326	zinc 1385	zinc 1514	zinc 1527	zinc 1531	zinc 17	zinc 18	zinc 189	zinc 249	zinc 252	zinc 28	zinc 32	zinc 370	zinc 38	zinc 41	zinc 447	zinc 48	zinc 52	zinc 53	zinc 530	zinc 57	zinc 58	zinc 660	zinc 7	zinc 707	zinc 888	zinc 948
zinc 10	25	1,00	0,20	0,40	0,20	0,33	0,75	0,33	0,00	0,00	0,20	0,40	0,20	0,20	0,20	0,60	0,20	0,20	0,20	0,50	0,40	0,60	0,00	0,40	0,20	0,40	0,40	0,00	0,60	0,60	0,20	0,40	0,20	0,33	0,20
zinc 1036	47	0,20	1,00	0,50	0,50	0,33	0,25	0,67	0,20	0,25	0,40	0,50	0,40	0,50	0,50	0,50	0,33	0,33	0,60	0,50	0,50	0,50	0,00	0,50	0,33	0,17	0,20	1,00	0,50	0,50	0,50	0,50	0,40	0,67	0,67
zinc 1037	43	0,40	0,50	1,00	0,25	0,33	0,50	0,67	0,20	0,25	0,40	1,00	0,40	0,50	0,71	0,67	0,29	0,33	0,60	0,50	0,63	0,60	0,00	0,67	0,29	0,29	0,40	1,00	0,36	0,40	0,67	0,42	0,60	0,67	0,43
zinc 11	43	0,20	0,50	0,25	1,00	0,33	0,25	0,67	0,20	0,25	0,40	0,25	0,40	0,67	0,43	0,44	0,43	0,67	0,60	0,50	0,50	0,60	0,00	0,44	0,71	0,29	0,20	1,00	0,55	0,50	0,50	0,67	0,40	0,67	0,43
zinc 1145	32	0,33	0,33	0,33	0,33	1,00	0,00	0,67	0,00	0,00	0,67	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,00	0,33	0,33	0,00	0,33	0,33	0,00	0,00	0,00	0,33	0,33	0,67	0,33	0,33	0,33	0,33
zinc 1146	31	0,75	0,25	0,50	0,25	0,00	1,00	0,33	0,25	0,25	0,25	0,50	0,25	0,25	0,25	0,75	0,25	0,25	0,25	1,00	0,50	0,75	0,00	0,50	0,25	0,75	0,75	1,00	0,75	0,75	0,25	0,50	0,00	0,00	0,25
zinc 1152	27	0,33	0,67	0,67	0,67	0,67	0,33	1,00	0,33	0,33	1,00	0,67	0,67	0,67	0,67	0,67	0,67	0,67	0,67	0,50	0,67	0,67	0,00	0,67	0,67	0,33	0,33	1,00	0,67	0,67	1,00	0,67	0,33	0,33	0,67
zinc 1326	29	0,00	0,20	0,20	0,20	0,00	0,25	0,33	1,00	1,00	0,40	0,20	0,80	0,40	0,20	0,20	0,20	0,20	0,20	0,50	0,20	0,20	0,00	0,20	0,20	0,20	0,20	1,00	0,20	0,20	0,40	0,20	0,00	0,00	0,20
zinc 1385	25	0,00	0,25	0,25	0,25	0,00	0,25	0,33	1,00	1,00	0,50	0,25	1,00	0,50	0,25	0,25	0,25	0,25	0,25	0,50	0,25	0,25	0,00	0,25	0,25	0,25	0,25	1,00	0,25	0,25	0,50	0,25	0,00	0,00	0,25
zinc 1514	40	0,20	0,40	0,40	0,40	0,67	0,25	1,00	0,40	0,50	1,00	0,40	0,60	0,60	0,40	0,40	0,40	0,40	0,40	0,50	0,40	0,40	0,00	0,40	0,40	0,20	0,20	1,00	0,40	0,40	0,80	0,40	0,20	0,33	0,40
zinc 1527	43	0,40	0,50	1,00	0,25	0,33	0,50	0,67	0,20	0,25	0,40	1,00	0,40	0,50	0,71	0,67	0,29	0,33	0,60	0,50	0,63	0,60	0,00	0,67	0,29	0,29	0,40	1,00	0,36	0,40	0,67	0,42	0,60	0,67	0,43
zinc 1531	28	0,20	0,40	0,40	0,40	0,33	0,25	0,67	0,80	1,00	0,60	0,40	1,00	0,60	0,40	0,40	0,40	0,40	0,40	0,50	0,40	0,40	0,00	0,40	0,40	0,20	0,20	1,00	0,40	0,40	0,60	0,40	0,20	0,33	0,40
zinc 17	34	0,20	0,50	0,50	0,67	0,33	0,25	0,67	0,40	0,50	0,60	0,50	0,60	1,00	0,50	0,67	0,50	0,50	0,60	0,50	0,67	0,67	0,00	0,67	0,50	0,33	0,20	1,00	0,67	0,67	0,67	0,67	0,40	0,67	0,50
zinc 18	30	0,20	0,50	0,71	0,43	0,33	0,25	0,67	0,20	0,25	0,40	0,71	0,40	0,50	1,00	0,71	0,29	0,33	0,60	0,50	0,57	0,71	0,00	0,71	0,43	0,14	0,20	1,00	0,43	0,43	0,67	0,71	0,60	0,67	0,43
zinc 189	44	0,60	0,50	0,67	0,44	0,33	0,75	0,67	0,20	0,25	0,40	0,67	0,40	0,67	0,71	1,00	0,43	0,50	0,60	1,00	0,88	0,89	0,00	0,78	0,43	0,43	0,40	1,00	0,67	0,67	0,67	0,67	0,60	0,67	0,43
zinc 249	38	0,20	0,33	0,29	0,43	0,33	0,25	0,67	0,20	0,25	0,40	0,29	0,40	0,50	0,29	0,43	1,00	0,50	0,60	0,50	0,43	0,43	0,00	0,43	0,43	0,29	0,20	1,00	0,43	0,43	0,33	0,43	0,20	0,33	0,29
zinc 252	45	0,20	0,33	0,33	0,67	0,33	0,25	0,67	0,20	0,25	0,40	0,33	0,40	0,50	0,33	0,50	0,50	1,00	0,60	0,50	0,50	0,67	0,00	0,50	0,67	0,33	0,20	1,00	0,67	0,67	0,33	0,83	0,20	0,33	0,33
zinc 28	26	0,20	0,60	0,60	0,60	0,33	0,25	0,67	0,20	0,25	0,40	0,60	0,40	0,60	0,60	0,60	0,60	0,60	1,00	0,50	0,60	0,60	0,00	0,60	0,40	0,20	0,20	1,00	0,60	0,60	0,60	0,80	0,40	0,67	0,60
zinc 32	22	0,50	0,50	0,50	0,50	0,00	1,00	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	1,00	0,50	0,50	0,50	1,00	0,50	1,00	0,00	0,50	0,50	0,50	0,50	1,00	1,00	1,00	0,50	0,50	0,00	0,00	0,50
zinc 370	41	0,40	0,50	0,63	0,50	0,33	0,50	0,67	0,20	0,25	0,40	0,63	0,40	0,67	0,57	0,88	0,43	0,50	0,60	0,50	1,00	0,75	0,00	0,88	0,43	0,43	0,60	1,00	0,63	0,63	0,67	0,75	0,60	0,67	0,43
zinc 38	39	0,60	0,50	0,60	0,60	0,33	0,75	0,67	0,20	0,25	0,40	0,60	0,40	0,67	0,71	0,89	0,43	0,67	0,60	1,00	0,75	1,00	0,00	0,78	0,71	0,43	0,40	1,00	0,80	0,70	0,67	0,80	0,60	0,67	0,43
zinc 41	22	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
zinc 447	45	0,40	0,50	0,67	0,44	0,33	0,50	0,67	0,20	0,25	0,40	0,67	0,40	0,67	0,71	0,78	0,43	0,50	0,60	0,50	0,88	0,78	0,00	1,00	0,43	0,43	0,60	1,00	0,56	0,56	0,67	0,67	0,60	0,67	0,43
zinc 48	30	0,20	0,33	0,29	0,71	0,33	0,25	0,67	0,20	0,25	0,40	0,29	0,40	0,50	0,43	0,43	0,43	0,67	0,40	0,50	0,43	0,71	0,00	0,43	1,00	0,29	0,20	1,00	0,71	0,57	0,33	0,86	0,20	0,33	0,29
zinc 52	35	0,40	0,17	0,29	0,29	0,00	0,75	0,33	0,20	0,25	0,20	0,29	0,20	0,33	0,14	0,43	0,29	0,33	0,20	0,50	0,43	0,43	0,00	0,43	0,29	1,00	0,60	1,00	0,43	0,43	0,17	0,43	0,00	0,00	0,14
zinc 53	20	0,40	0,20	0,40	0,20	0,00	0,75	0,33	0,20	0,25	0,20	0,40	0,20	0,20	0,20	0,40	0,20	0,20	0,20	0,50	0,60	0,40	0,00	0,60	0,20	0,60	1,00	1,00	0,40	0,40	0,20	0,40	0,00	0,00	0,20
zinc 530	34	0,00	1,00	1,00	1,00	0,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00	0,00	1,00
zinc 57	48	0,60	0,50	0,36	0,55	0,33	0,75	0,67	0,20	0,25	0,40	0,36	0,40	0,67	0,43	0,67	0,43	0,67	0,60	1,00	0,63	0,80	0,00	0,56	0,71	0,43	0,40	1,00	1,00	1,00	0,50	0,64	0,40	0,67	0,43
zinc 58	45	0,60	0,50	0,40	0,50	0,33	0,75	0,67	0,20	0,25	0,40	0,40	0,40	0,67	0,43	0,67	0,43	0,67	0,60	1,00	0,63	0,70	0,00	0,56	0,57	0,43	0,40	1,00	1,00	1,00	0,50	0,60	0,40	0,67	0,43
zinc 660	32	0,20	0,50	0,67	0,50	0,67	0,25	1,00	0,40	0,50	0,80	0,67	0,60	0,67	0,67	0,67	0,33	0,33	0,60	0,50	0,67	0,67	0,00	0,67	0,33	0,17	0,20	1,00	0,50	0,50	1,00	0,67	0,60	0,67	0,50
zinc 7	43	0,40	0,50	0,42	0,67	0,33	0,50	0,67	0,20	0,25	0,40	0,42	0,40	0,67	0,71	0,67	0,43	0,83	0,80	0,50	0,75	0,80	0,00	0,67	0,86	0,43	0,40	1,00	0,64	0,60	0,67	1,00	0,60	0,67	0,43
zinc 707	36	0,20	0,40	0,60	0,40	0,33	0,00	0,33	0,00	0,00	0,20	0,60	0,20	0,40	0,60	0,60	0,20	0,20	0,40	0,00	0,60	0,60	0,00	0,60	0,20	0,00	0,00	0,00	0,40	0,40	0,60	0,60	1,00	0,67	0,40
zinc 888	37	0,33	0,67	0,67	0,67	0,33	0,00	0,33	0,00	0,00	0,33	0,67	0,33	0,67	0,67	0,67	0,33	0,33	0,67	0,00	0,67	0,67	0,00	0,67	0,33	0,00	0,00	0,00	0,67	0,67	0,67	0,67	0,67	1,00	0,67
zinc 948	51	0,20	0,67	0,43	0,43	0,33	0,25	0,67	0,20	0,25	0,40	0,43	0,40	0,50	0,43	0,43	0,29	0,33	0,60	0,50	0,43	0,43	0,00	0,43	0,29	0,14	0,20	1,00	0,43	0,43	0,50	0,43	0,40	0,67	1,00

COSINE	atom	zinc 10	zinc 1036	zinc 1037	zinc 11	zinc 1145	zinc 1146	zinc 1152	zinc 1326	zinc 1385	zinc 1514	zinc 1527	zinc 1531	zinc 17	zinc 18	zinc 189	zinc 249	zinc 252	zinc 28	zinc 32	zinc 370	zinc 38	zinc 41	zinc 447	zinc 48	zinc 52	zinc 53	zinc 530	zinc 57	zinc 58	zinc 660	zinc 7	zinc 707	zinc 888	zinc 948
zinc 10	25	1,00	0,18	0,26	0,12	0,26	0,67	0,26	0,00	0,00	0,20	0,26	0,20	0,18	0,17	0,45	0,17	0,18	0,20	0,32	0,32	0,42	0,00	0,30	0,17	0,34	0,40	0,00	0,40	0,42	0,18	0,26	0,20	0,26	0,17
zinc 1036	47	0,18	1,00	0,35	0,34	0,24	0,20	0,47	0,18	0,20	0,37	0,35	0,37	0,50	0,46	0,41	0,31	0,33	0,55	0,29	0,43	0,39	0,00	0,41	0,31	0,15	0,18	0,41	0,37	0,39	0,50	0,35	0,37	0,47	0,62
zinc 1037	43	0,26	0,35	1,00	0,24	0,17	0,29	0,33	0,13	0,14	0,26	1,00	0,26	0,35	0,55	0,58	0,22	0,24	0,39	0,20	0,51	0,55	0,00	0,58	0,22	0,22	0,26	0,29	0,35	0,37	0,47	0,42	0,39	0,33	0,33
zinc 11	43	0,12	0,34	0,24	1,00	0,16	0,14	0,32	0,12	0,14	0,25	0,24	0,25	0,45	0,31	0,37	0,31	0,45	0,37	0,20	0,39	0,53	0,00	0,37	0,52	0,21	0,12	0,28	0,50	0,44	0,34	0,64	0,25	0,32	0,31
zinc 1145	32	0,26	0,24	0,17	0,16	1,00	0,00	0,67	0,00	0,00	0,52	0,17	0,26	0,24	0,22	0,19	0,22	0,24	0,26	0,00	0,20	0,18	0,00	0,19	0,22	0,00	0,00	0,00	0,17	0,18	0,47	0,17	0,26	0,33	0,22
zinc 1146	31	0,67	0,20	0,29	0,14	0,00	1,00	0,29	0,22	0,25	0,22	0,29	0,22	0,20	0,19	0,50	0,19	0,20	0,22	0,71	0,35	0,47	0,00	0,33	0,19	0,57	0,67	0,50	0,45	0,47	0,20	0,29	0,00	0,00	0,19
zinc 1152	27	0,26	0,47	0,33	0,32	0,67	0,29	1,00	0,26	0,29	0,77	0,33	0,52	0,47	0,44	0,38	0,44	0,47	0,52	0,41	0,41	0,37	0,00	0,38	0,44	0,22	0,26	0,58	0,35	0,37	0,71	0,33	0,26	0,33	0,44
zinc 1326	29	0,00	0,18	0,13	0,12	0,00	0,22	0,26	1,00	0,89	0,40	0,13	0,80	0,37	0,17	0,15	0,17	0,18	0,20	0,32	0,16	0,14	0,00	0,15	0,17	0,17	0,20	0,45	0,13	0,14	0,37	0,13	0,00	0,00	0,17
zinc 1385	25	0,00	0,20	0,14	0,14	0,00	0,25	0,29	0,89	1,00	0,45	0,14	0,89	0,41	0,19	0,17	0,19	0,20	0,22	0,35	0,18	0,16	0,00	0,17	0,19	0,19	0,22	0,50	0,15	0,16	0,41	0,14	0,00	0,00	0,19
zinc 1514	40	0,20	0,37	0,26	0,25	0,52	0,22	0,77	0,40	0,45	1,00	0,26	0,60	0,55	0,34	0,30	0,34	0,37	0,40	0,32	0,32	0,28	0,00	0,30	0,34	0,17	0,20	0,45	0,27	0,28	0,73	0,26	0,20	0,26	0,34
zinc 1527	43	0,26	0,35	1,00	0,24	0,17	0,29	0,33	0,13	0,14	0,26	1,00	0,26	0,35	0,55	0,58	0,22	0,24	0,39	0,20	0,51	0,55	0,00	0,58	0,22	0,22	0,26	0,29	0,35	0,37	0,47	0,42	0,39	0,33	0,33
zinc 1531	28	0,20	0,37	0,26	0,25	0,26	0,22	0,52	0,80	0,89	0,60	0,26	1,00	0,55	0,34	0,30	0,34	0,37	0,40	0,32	0,32	0,28	0,00	0,30	0,34	0,17	0,20	0,45	0,27	0,28	0,55	0,26	0,20	0,26	0,34
zinc 17	34	0,18	0,50	0,35	0,45	0,24	0,20	0,47	0,37	0,41	0,55	0,35	0,55	1,00	0,46	0,54	0,46	0,50	0,55	0,29	0,58	0,52	0,00	0,54	0,46	0,31	0,18	0,41	0,49	0,52	0,67	0,47	0,37	0,47	0,46
zinc 18	30	0,17	0,46	0,55	0,31	0,22	0,19	0,44	0,17	0,19	0,34	0,55	0,34	0,46	1,00	0,63	0,29	0,31	0,51	0,27	0,53	0,60	0,00	0,63	0,43	0,14	0,17	0,38	0,34	0,36	0,62	0,55	0,51	0,44	0,43
zinc 189	44	0,45	0,41	0,58	0,37	0,19	0,50	0,38	0,15	0,17	0,30	0,58	0,30	0,54	0,63	1,00	0,38	0,41	0,45	0,47	0,82	0,84	0,00	0,78	0,38	0,38	0,30	0,33	0,60	0,63	0,54	0,58	0,45	0,38	0,38
zinc 249	38	0,17	0,31	0,22	0,31	0,22	0,19	0,44	0,17	0,19	0,34	0,22	0,34	0,46	0,29	0,38	1,00	0,46	0,51	0,27	0,40	0,36	0,00	0,38	0,43	0,29	0,17	0,38	0,34	0,36	0,31	0,33	0,17	0,22	0,29
zinc 252	45	0,18	0,33	0,24	0,45	0,24	0,20	0,47	0,18	0,20	0,37	0,24	0,37	0,50	0,31	0,41	0,46	1,00	0,55	0,29	0,43	0,52	0,00	0,41	0,62	0,31	0,18	0,41	0,49	0,52	0,33	0,59	0,18	0,24	0,31
zinc 28	26	0,20	0,55	0,39	0,37	0,26	0,22	0,52	0,20	0,22	0,40	0,39	0,40	0,55	0,51	0,45	0,51	0,55	1,00	0,32	0,47	0,42	0,00	0,45	0,34	0,17	0,20	0,45	0,40	0,42	0,55	0,52	0,40	0,52	0,51
zinc 32	22	0,32	0,29	0,20	0,20	0,00	0,71	0,41	0,32	0,35	0,32	0,20	0,32	0,29	0,27	0,47	0,27	0,29	0,32	1,00	0,25	0,45	0,00	0,24	0,27	0,27	0,32	0,71	0,43	0,45	0,29	0,20	0,00	0,00	0,27
zinc 370	41	0,32	0,43	0,51	0,39	0,20	0,35	0,41	0,16	0,18	0,32	0,51	0,32	0,58	0,53	0,82	0,40	0,43	0,47	0,25	1,00	0,67	0,00	0,82	0,40	0,40	0,47	0,35	0,53	0,56	0,58	0,61	0,47	0,41	0,40
zinc 38	39	0,42	0,39	0,55	0,53	0,18	0,47	0,37	0,14	0,16	0,28	0,55	0,28	0,52	0,60	0,84	0,36	0,52	0,42	0,45	0,67	1,00	0,00	0,74	0,60	0,36	0,28	0,32	0,76	0,70	0,52	0,73	0,42	0,37	0,36
zinc 41	22	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
zinc 447	45	0,30	0,41	0,58	0,37	0,19	0,33	0,38	0,15	0,17	0,30	0,58	0,30	0,54	0,63	0,78	0,38	0,41	0,45	0,24	0,82	0,74	0,00	1,00	0,38	0,38	0,45	0,33	0,50	0,53	0,54	0,58	0,45	0,38	0,38
zinc 48	30	0,17	0,31	0,22	0,52	0,22	0,19	0,44	0,17	0,19	0,34	0,22	0,34	0,46	0,43	0,38	0,43	0,62	0,34	0,27	0,40	0,60	0,00	0,38	1,00	0,29	0,17	0,38	0,57	0,48	0,31	0,65	0,17	0,22	0,29
zinc 52	35	0,34	0,15	0,22	0,21	0,00	0,57	0,22	0,17	0,19	0,17	0,22	0,17	0,31	0,14	0,38	0,29	0,31	0,17	0,27	0,40	0,36	0,00	0,38	0,29	1,00	0,51	0,38	0,34	0,36	0,15	0,33	0,00	0,00	0,14
zinc 53	20	0,40	0,18	0,26	0,12	0,00	0,67	0,26	0,20	0,22	0,20	0,26	0,20	0,18	0,17	0,30	0,17	0,18	0,20	0,32	0,47	0,28	0,00	0,45	0,17	0,51	1,00	0,45	0,27	0,28	0,18	0,26	0,00	0,00	0,17
zinc 530	34	0,00	0,41	0,29	0,28	0,00	0,50	0,58	0,45	0,50	0,45	0,29	0,45	0,41	0,38	0,33	0,38	0,41	0,45	0,71	0,35	0,32	0,00	0,33	0,38	0,38	0,45	1,00	0,30	0,32	0,41	0,29	0,00	0,00	0,38
zinc 57	48	0,40	0,37	0,35	0,50	0,17	0,45	0,35	0,13	0,15	0,27	0,35	0,27	0,49	0,34	0,60	0,34	0,49	0,40	0,43	0,53	0,76	0,00	0,50	0,57	0,34	0,27	0,30	1,00	0,95	0,37	0,61	0,27	0,35	0,34
zinc 58	45	0,42	0,39	0,37	0,44	0,18	0,47	0,37	0,14	0,16	0,28	0,37	0,28	0,52	0,36	0,63	0,36	0,52	0,42	0,45	0,56	0,70	0,00	0,53	0,48	0,36	0,28	0,32	0,95	1,00	0,39	0,55	0,28	0,37	0,36
zinc 660	32	0,18	0,50	0,47	0,34	0,47	0,20	0,71	0,37	0,41	0,73	0,47	0,55	0,67	0,62	0,54	0,31	0,33	0,55	0,29	0,58	0,52	0,00	0,54	0,31	0,15	0,18	0,41	0,37	0,39	1,00	0,47	0,55	0,47	0,46
zinc 7	43	0,26	0,35	0,42	0,64	0,17	0,29	0,33	0,13	0,14	0,26	0,42	0,26	0,47	0,55	0,58	0,33	0,59	0,52	0,20	0,61	0,73	0,00	0,58	0,65	0,33	0,26	0,29	0,61	0,55	0,47	1,00	0,39	0,33	0,33
zinc 707	36	0,20	0,37	0,39	0,25	0,26	0,00	0,26	0,00	0,00	0,20	0,39	0,20	0,37	0,51	0,45	0,17	0,18	0,40	0,00	0,47	0,42	0,00	0,45	0,17	0,00	0,00	0,00	0,27	0,28	0,55	0,39	1,00	0,52	0,34
zinc 888	37	0,26	0,47	0,33	0,32	0,33	0,00	0,33	0,00	0,00	0,26	0,33	0,26	0,47	0,44	0,38	0,22	0,24	0,52	0,00	0,41	0,37	0,00	0,38	0,22	0,00	0,00	0,00	0,35	0,37	0,47	0,33	0,52	1,00	0,44
zinc 948	51	0,17	0,62	0,33	0,31	0,22	0,19	0,44	0,17	0,19	0,34	0,33	0,34	0,46	0,43	0,38	0,29	0,31	0,51	0,27	0,40	0,36	0,00	0,38	0,29	0,14	0,17	0,38	0,34	0,36	0,46	0,33	0,34	0,44	1,00